United States General Accounting Office

GAO

Report to the Chairman and Ranking Minority Member, Subcommittee on Readiness and Management Support, Committee on Armed Services, U.S. Senate

July 2000

BEST PRACTICES

A More Constructive Test Approach Is Key to Better Weapon System Outcomes

G A O

Accountability · Integrity · Reliability

GAO/NSIAD-00-199

Contents

Abbreviations

DOD	Department of Defense
THAAD	Theater High Altitude Area Defense
SLAM-ER	Stand-off Land Attack Missile-Expanded Response

GAO
Accountability * Integrity * Reliability

United States General Accounting Office
Washington, D.C. 20548

National Security and
International Affairs Division

B-282345

July 31, 2000

The Honorable James Inhofe
Chairman
The Honorable Charles Robb
Ranking Minority Member
Subcommittee on Readiness and Management Support
Committee on Armed Services
United States Senate

As you requested, this report addresses how best commercial practices for testing and evaluating new products offer ways to improve the way the Department of Defense conducts test and evaluation on weapon systems. It also addresses how differences in the commercial and the Department environments for testing and evaluating new products affect the corresponding practices.

We are sending copies of this report to other congressional committees; the Secretaries of Defense, the Army, the Navy, and the Air Force; and the Director, Office of Management and Budget.

If you or your staff have any questions, I can be reached at (202) 512-4841. Contacts and key contributors to this report are listed in appendix III.

Katherine V. Schinasi
Associate Director
Defense Acquisition Issues

Executive Summary

Purpose

Despite good intentions and some progress by the Department of Defense (DOD), weapon system programs still suffer from persistent problems associated with late or incomplete testing. Often, the fate of a program is jeopardized by unexpectedly poor test results. In such cases, testing becomes a watershed event that attracts unwanted attention from decisionmakers and critics. The discovery of problems in complex products is a normal part of any development process, and testing is perhaps the most effective tool for discovering such problems. However, why surprises in testing repeatedly occur and why such results polarize organizations into proponents and critics of programs have proven elusive questions to answer. Indeed, numerous solutions proposed over the years by different DOD leaders and distinguished outside panels have not had much effect.

Lessons learned by leading commercial firms in developing new products are applicable to the management and testing of weapon systems. These firms achieve the type of outcomes DOD seeks: they develop more sophisticated products faster and less expensively than their predecessors. Commercial firms have found constructive ways of conducting testing and evaluation that help them avoid being surprised by problems late in a product's development. In response to a request from the Chairman and the Ranking Minority Member, Subcommittee on Readiness and Management Support, Senate Committee on Armed Services, GAO examined (1) how the conduct of testing and evaluation affects commercial and DOD program outcomes, (2) how best commercial testing and evaluation practices compare with DOD's, and (3) what factors account for the differences in these practices.

Background

The fundamental purpose of testing and evaluation does not differ for military and commercial products. Testing is the main instrument used to gauge the progress being made when an idea or concept is translated into an actual product. Evaluation refers to what is learned from a test. Testing and evaluation is used at a variety of levels, including basic technology, components and subsystems, and a complete system or product. The ultimate goal of testing and evaluation is to make sure the product works as intended before it is provided to customers. In both DOD and commercial firms, product testing is conducted by organizations separate from those responsible for managing product development.

Among the key sources of information GAO relied on for this report were individual DOD acquisition programs and commercial firms, including Boeing, Intel, Dupont, AT&T, and General Electric. These firms are recognized as leaders in developing high-quality products on time and within budget. In this report, GAO highlights these firms' practices in testing and evaluating new products. As such, these practices are not intended to describe those of all commercial industry or suggest that commercial firms are without fault.

Results in Brief

For the leading commercial firms GAO visited, the proof of testing and evaluation lies in whether a product experiences what one firm called "late-cycle churn," or the scramble to fix a significant problem discovered late in development. Nearly all the firms had experienced such problems on some of their previous products but used testing and evaluation to preclude such problems on new products. Late cycle churn has been a fairly common occurrence on DOD weapon systems. Often, tests of a full system, such as launching a missile, identify problems that could have been found earlier. Typically, DOD's response to such test results is to expend more time and money to solve the problems. Only rarely are programs terminated. Problems revealed in flight tests caused two programs GAO reviewed—the Theater High Altitude Area Defense system and the DarkStar unmanned aerial vehicle[1]—to take twice as long to develop as planned.

The leading commercial firms GAO visited use testing and other techniques to expose problems earlier than the DOD programs GAO reviewed. The firms focus on validating that their products have reached increasing levels of product maturity at given points in time. The firms' products have three maturity levels in common: components work individually, components work together as a system in a controlled setting, and components work together as a system in a realistic setting. The key to minimizing late surprises is to reach the first two levels early, limiting the burden on the third level. By concentrating on validating knowledge rather than the specific technique used—such as testing—commercial firms avoid skipping key events and holding hollow tests that do not add knowledge.

[1]The Theater High Altitude Area Defense weapon is a mobile, ground-based missile system designed to hit and destroy incoming ballistic missiles. The DarkStar unmanned aerial vehicle program was designed to provide theater reconnaissance and surveillance to operational commanders.

On the weapon programs, system level testing carried a greater share of the burden. Earlier tests were delayed, skipped, or not conducted in a way that advanced knowledge. For example, several failures in flight tests of the Theater High Altitude Area Defense system were traced to problems that could have been discovered in ground testing.

The differences in testing practices reflect the different demands commercial firms and DOD impose on program managers. Leading commercial firms have learned to insist that a product satisfy the customer and make a profit. Success is threatened if managers are unduly optimistic or if unknowns about a product are not resolved early, when costs are low and more options are available. The role of testing under these circumstances is constructive, for it helps eliminate unknowns. Product managers view testers and realistic test plans as contributing to a product's success. Success for a weapon system program is different; it centers on attempting to provide a superior capability within perceived time and funding limits. Success is influenced by the competition for funding and the quest for top performance; delivering the product late and over cost does not necessarily threaten success. Testing plays a less constructive role in DOD because a failure in a key test can jeopardize program support. Specifically, test results often become directly linked to funding and other key decisions for programs. Such a role creates a more adversarial relationship between testers and program managers.

Principal Findings

Problems Found Late in Development Signal Weaknesses in Testing and Evaluation

Over the years, GAO found numerous examples of late-cycle churn in DOD programs, regardless of their size, complexity, or product type. More recent examples include the following:

- The DarkStar unmanned aerial vehicle crashed during initial flight tests. DOD spent twice the planned money and time to redesign and retest the aircraft, eventually terminating the program.
- The Theater High Altitude Area Defense missile program was nearly terminated after eight consecutive flight test failures. Instead of taking 4 years, the Army spent 8 years developing the missile. The last two flight tests in 1999 were successful.
- The Army bought 6,700 cargo trailers before tests revealed that the trailers damaged the trucks they were hitched to. As a result, the trailers

required extensive modifications. The majority of the trailers are currently in storage.

Commercial firms have learned from making similar mistakes. For example, Boeing experienced significant problems with the 747-400 airliner; the problems caused the company to deliver the aircraft late and to assign 300 engineers to solve problems not found earlier in development. Its testing approach was so much more effective on the 777-200 airliner that Boeing reduced change, error, and rework by more than 60 percent. In addition, the Federal Aviation Administration certified the initial aircraft for overseas flight on the basis of test results. The certification normally requires 2 years of actual flight service. After a flaw in the original Pentium® microprocessor cost Intel about $500 million to replace products for customers, the firm approached the testing of subsequent microprocessors differently. The quality of these microprocessors, such as the Pentium® Pro and Pentium® III, has significantly improved, yet they were developed in the same amount of time as the original Pentium® microprocessor, despite being many times more complex.

Testing Early to Validate Product Knowledge Is a Best Practice

Leading commercial firms GAO visited think in terms of validating that a product works as intended and use testing and evaluation as a means to that end. To limit the burden on the product's third maturity level (operating in a realistic environment), leading firms ensure that (1) the right validation events—tests, simulations, and other means for demonstrating product maturity—occur at the right times, (2) each validation event produces quality results, and (3) the knowledge gained from an event is used to improve the product. The firms hold challenging tests early to expose weaknesses in a product's design. AT&T refers to this as a "break it big early" philosophy. To reduce the burden on later testing, Boeing made extensive investments in computer-aided design techniques and a system integration laboratory that could test all of the 777-200's main components in simulated flight conditions. Intel's most significant improvement in validation has been in the design stage—before any prototype microprocessors are made. Its revamped validation techniques have enabled testers to identify most flaws before a prototype is made and have reduced the number of prototype iterations needed.

The weapon system programs GAO reviewed had a greater tendency to attempt to reach all three product maturity levels in one step in the late stages of development. For example, knowledge typically gained during component testing was not validated before flight testing began on the Theater High Altitude Area Air Defense system. Many components, like the seeker, which finds and tracks the intended target, were shipped for flight tests without having been ground tested; they later contributed to flight test failures. Validation of system level maturity was limited on the Navy's Standoff Land Attack Missile-Extended Range[2] for different reasons. Although the program followed a disciplined development process, with over 6,000 tests, problems experienced in a predecessor missile were excluded. Also, conditions for system level tests were not realistic, which lowered the value of the information gained and masked some missile limitations. These limitations contributed to the missile's failure when the customer used the system in realistic test conditions.

Different Incentives Make Testing a More Constructive Factor in Commercial Programs Than in Weapon System Programs

Leading commercial firms GAO visited adopted best practices because they gained a better appreciation for why testing is done versus how it is done. Full corporate support for new product developments defuses test results as a threat to program support and enables testers to contribute throughout product development. Candor is rewarded by a product's success. The manager of Boeing's 777-200 program viewed test problems as "gems to be mined" and stressed that the earlier a problem is discovered, the less expensive it is to fix. DuPont has undergone a similar cultural change. A test failure used to mean that a product did not meet expectations; now, DuPont sees a test failure as meaning that knowledge was not gained. Intel has succeeded in getting its validation staff to actively seek out and communicate problems to product managers to improve a product's success. The role testers have in a commercial product is not determined by their organizational position or ability to withhold approval; it is because (1) they help a product succeed and (2) they are credible and have earned the confidence of product developers.

GAO's previous and current work has shown that it is difficult for a weapon system program to compete for approval unless it offers significantly better performance than other weapons, yet fits within available funding and planned schedules. There are thus greater incentives for managers to

[2]The SLAM-ER is a Navy missile, which will be used on aircraft carriers and launched from F/A-18 aircraft to make precision strikes against land targets.

accept immature technologies and make optimistic assessments about what can be accomplished with limited resources. Test results tend to become scorecards that demonstrate whether the program is ready to proceed or to receive the next increment of funding. Whereas testing and evaluation of commercial products mainly benefits the product manager, in DOD, testing and evaluation is more for the benefit of the testers and decisionmakers above the program manager. Managers thus have incentives to postpone difficult tests and to limit open communication about test results. Externally imposed constraints on cost or schedule can intensify these incentives. Pressures to meet an early fielding date caused managers of the Theater High Altitude Area Air Defense system to cut back efforts to validate the first two product maturity levels and to overrule the objections of testers. Managers in both the DarkStar unmanned aerial vehicle and the Standoff Land Attack Missile programs also overruled testers because of funding and schedule pressures.

Recommendations

To lessen the dependence on testing late in development and to foster a more constructive relationship between program managers and testers, GAO recommends that the Secretary of Defense instruct acquisition managers to structure test plans around the attainment of increasing levels of product maturity, orchestrate the right mix of tools to validate these maturity levels, and build and resource acquisition strategies around this approach. GAO also recommends that validation of lower levels of product maturity not be deferred to the third level. Finally, GAO recommends that the Secretary require that weapon systems demonstrate a specified level of product maturity before major programmatic approvals.

Agency Comments

DOD committed to establishing appropriate levels of product maturity, and agreed with two of the three recommendations. It disagreed with GAO's third recommendation, which originally called for DOD not to schedule major test events in the same budget year as major programmatic or funding decisions. DOD stated that the recommendation would delay the delivery of weapon systems and increase costs. GAO has reworded the recommendation, dropping the language on holding major test events and program decisions in different years, and substituting the language on demonstrating product maturity before major programmatic approvals. A discussion of DOD's comments appears in chapter 5, and the comments appear in full in appendix I.

Introduction

Someone new to the study of weapon systems might observe the turmoil around problems that testing revealed in new weapons like the Army's cargo trailer and wonder why they were not found and corrected earlier. A more seasoned observer will recognize this turmoil as a replay of what has often happened in past programs, such as the C-17 Airlifter and the Sergeant York Air Defense Gun. In such cases, testing becomes a watershed event in the weapon's survival and attracts much unwanted attention from decisionmakers and critics. Sometimes, test results prompt the cancellation of a program after the bulk of the development investment has been made, as with Sergeant York. In other cases, like the C-17, substantial schedule and cost increases are accepted to redesign the weapon system. In still others, key tests are completed after production has begun—or, in the case of the B-1B bomber, after production is completed—necessitating very costly retrofits.

The discovery of problems in complex products is a normal part of any development process, and testing is perhaps the most effective tool for discovering such problems. However, why problems in Department of Defense (DOD) testing repeatedly occur and why test results polarize organizations into proponents and critics of programs have proven elusive questions to answer. Indeed, numerous solutions proposed over the years by different DOD leaders and distinguished outside panels, as well as reorganizations within the Department, have not made much difference in the test experience of weapon systems.

Our previous work has disclosed that the lessons learned by leading commercial firms in different aspects of product development, such as the maturation of new technologies and the building of good business relationships with suppliers, are applicable to the management of weapon systems. These firms are developing new products with the types of outcomes DOD seeks: more sophisticated designs than their predecessors but developed faster and less expensively. While leading commercial companies employ testing techniques and tools that are similar to DOD, they have found ways to apply these tools and techniques with more constructive results. Testing and evaluation have become important ingredients to the firms' ability to obtain better outcomes for newly developed products. This approach holds promise for DOD. On the other hand, proceeding under current testing and evaluation practices will continue to disclose serious problems in the late stages of development, when the cost to correct them is very high.

The Role of Testing and Evaluation in Product Development

The fundamental purpose of testing does not differ for military and commercial products. Testing is perhaps the main instrument used to gauge the progress being made when an idea or concept is translated into an actual product that people use. Evaluation refers to the analysis of the meaning of test results and what can be learned from them. Ideally, testing progresses from early laboratory testing of technologies, to component and subsystem testing, through testing of a complete system, and finally to trial use in the customer's hands. To be of value at each stage, test results must be credible and used to improve the product. If a test has been poorly designed or has not been properly controlled, its results may not be usable. On the other hand, if test results are credible but are not properly evaluated or used, they do not help the product mature.

To manage its testing process, DOD has developed a complex organization that includes acquisition, test, and oversight officials in the services and in the Office of the Secretary of Defense. In addition, individual weapon systems are subject to specific congressional direction regarding the conduct of their test programs. For example, Congress specifically directed[1] that the Secretary of Defense certify that the F-22 fighter aircraft program had completed 433 hours (about 10 percent of the planned flight test hours) before it began production. If this level of testing was not achieved, the Secretary was required to justify to Congress the reasons why.

DOD divides testing into two categories: developmental and operational. The goal of developmental tests is to determine whether the weapon system meets the technical specifications of the contract. Developmental testing is done by contractors, university and government labs, and various organizations within each military service. The goal of operational testing is to evaluate the effectiveness and suitability of the weapon system in realistic combat conditions. Operational testing is managed by different military test organizations that represent the customers, such as the combat units that will use the weapons. Each service has its own operational test organization and associated test ranges. Operational testers have more independence than developmental testers; they provide their results to Congress as well as to senior officials in the services and the Office of the Secretary of Defense.

[1]P.L. 105-261, section 131.

Congress has been particularly interested in operational testing. In 1983, Congress established the office of the Director of Operational Test and Evaluation to effect several reforms concerning operational testing. Prominent among the reform objectives were independent oversight and coordination of the military services' planning and execution of operational tests, and objective reporting of those results to decisionmakers in DOD and Congress.

Leading commercial firms also have organizations dedicated to testing new products, but these organizations are more integrated with product managers. Commercial firms generally do not make a distinction between developmental and operational testing. One reason for this is that new technologies are aggressively tested and well understood before commercial firms allow them in a new product development. Another reason they do not single out developmental testing is that they are developing the product themselves—not receiving it through a contract. Unlike DOD, commercial firms are not typically subject to specific congressional direction regarding their test programs and therefore have more freedom to develop and test products without external restrictions. However, many firms are subject to some regulatory oversight by other government agencies such as the National Transportation Safety Board and the Federal Aviation Administration (for commercial aircraft) and the Food and Drug Administration (for chemicals used in commercial products).

Testing is done by leading commercial firms within the broader context of a knowledge-based product development process. In an earlier report, we described this process as having three key junctures, or knowledge points.[2] These are

- knowledge point 1: when a match is made between the customer's requirements and the available technology;
- knowledge point 2: when the product's design is determined to be capable of meeting performance requirements; and
- knowledge point 3: when the product is determined to be producible within cost, schedule, and quality targets.

[2]*Best Practices: Successful Application to Weapon Acquisitions Requires Changes in DOD's Environment* (GAO/NSIAD-98-56, Feb. 24, 1998).

Problems in DOD Testing and Evaluation Remain, Despite Numerous Reforms

Many studies over the years have acknowledged problems with DOD's acquisition approach, including testing, and have attempted to reform or improve the process. DOD itself has recognized the need to reform and has tried a variety of approaches to this end—streamlining acquisition organizations, mandating career and training requirements for its workforce, and establishing independent test organizations in each service—with limited success. In the 1970s, DOD adopted a "fly before buy" policy to ensure that weapon systems were more thoroughly tested prior to committing to a production decision. In 1981, the Deputy Secretary of Defense, Frank Carlucci, noted weaknesses in testing and recommended initiatives to increase test hardware so that the designing and testing of subsystems, systems, and software could be conducted thoroughly and efficiently. Five years later, the Packard Commission recommended improvements to early prototype testing.[3] More recently, a 1999 Defense Science Board study concluded that testing must be addressed earlier in the development process.[4] It advocated that operational test personnel should be involved in the early acquisition stages to provide critical testing perspectives to acquisition planners. In addition, a 1999 Science Applications International Corporation report cautioned that although DOD's goal of reducing cycle time for weapon system development and production was valid, curtailing testing was not an option because it was already at a minimum level.[5]

Despite good intentions and some progress, DOD weapon programs still suffer from persistent problems associated with late or incomplete testing. Many weapons still begin production with only a minimal amount of knowledge gained through testing. Our ongoing reviews of DOD's major weapon system acquisitions show that significant reforms have not yet been reflected in the management of and decision-making for individual programs. Over the years, we have reported on testing issues, such as unexpected performance problems, inadequate component testing, difficulties with software/hardware integration, deletion of test events, and limited analysis of test results. Such problems occur regardless of weapons' complexity or the era in which they were procured. Invariably,

[3]A Quest For Excellence: Final Report to the President by the President's Blue Ribbon Commission on Defense Management, June 1986.

[4]Report of the Defense Science Board on Test and Evaluation, September 1999.

[5]Best Practices Applicable to DOD Developmental Test and Evaluation, June 1999.

test weaknesses cause negative program outcomes, such as cost increases, schedule delays, or performance shortfalls.

Objectives, Scope, and Methodology

The Chairman and the Ranking Minority Member, Subcommittee on Readiness and Management Support, Senate Committee on Armed Services, requested that we examine various aspects of the acquisition process to identify best practices that can improve the outcomes of weapon system programs. To date, we have issued reports on advanced quality concepts, earned value management techniques used to assess progress of research and development contracts, management of a product's transition from development to production, management of the supplier base, technology maturation, and training for best practices (see related GAO products). This report covers the best practices for testing and evaluating new products. Our overall objective was to evaluate whether best practices in testing and evaluation offer methods or strategies that could improve the way DOD manages weapon systems. Specifically, we examined (1) how the conduct of testing and evaluation affects commercial and DOD program outcomes, (2) how best commercial testing and evaluation practices compare with DOD's, and (3) what factors account for differences in testing practices.

To obtain the above information and identify the best testing practices in the commercial sector, we conducted literature searches and contacted universities, industry associations, testing laboratories, and experts and consultants in the area of testing for new development products. On the basis of these discussions and analyses, we selected several world-class companies with a solid track record for developing high-quality products. We used structured interview questions sent in advance of our visits to gather uniform information about each firm's testing practices and the results achieved. After our visits, we analyzed data from each company and identified best testing practices used by these firms. We then prepared and distributed a depiction of these practices to each firm we contacted. We incorporated their comments and insights in our subsequent analyses; we also provided each firm a copy of our draft report for review and comment.

We did not attempt to select only those commercial firms whose products have the most in common with weapon systems. Such an approach would have limited our ability to obtain an understanding of best practices in testing from a diverse group of recognized industry leaders. The firms we selected represent markedly different industry sectors and product lines.

Nevertheless, the testing practices and approaches exhibited similarities. The firms we visited were

- AT&T, Warrenville, Illinois,
- Boeing Company, Seattle, Washington,
- DuPont, Inc. Wilmington, Delaware,
- General Electric Aircraft Engines, Evendale, Ohio, and
- INTEL, Hillsboro, Oregon.

Our report summarizes a number of the best commercial practices in testing and evaluation. We did not intend to describe all commercial industry practices or suggest that all commercial firms continually use best practices. Also, we were limited in our ability to obtain and present some relevant data that commercial firms considered proprietary in nature. Due to the highly competitive nature of their businesses, the firms did not wish to release specific details of how their current product lines achieved successful test outcomes.

To better understand DOD's testing and evaluation practices, we reviewed current DOD and service policy directives and guidance on testing and evaluation. We met with officials from the Director of Operational Testing and Evaluation, Deputy for Developmental Testing and Evaluation in the Office of the Secretary of Defense, and test officials from Army and Air Force headquarters. The Navy provided written responses to our questions. We analyzed recent studies of DOD testing and evaluation by external organizations such as the Defense Science Board and Science Applications International Corporation. We also conducted detailed work on four individual weapon programs: the Theater High Altitude Area Defense (THAAD) missile, the DarkStar unmanned aerial vehicle, the Standoff Land Attack Missile-Expanded Response (SLAM-ER), and the F-22 Raptor aircraft. We also examined information from GAO and DOD Inspector General reports on the testing experiences of other weapon systems.

The THAAD is a mobile ground-based missile system designed to hit and destroy incoming ballistic missiles. It is jointly managed by the Army and the Ballistic Missile Defense Organization. THAAD is expected to provide higher altitude missile defense in concert with lower altitude systems like the Patriot missile system. It consists of mobile launchers; interceptors; radars; battle management/command, control, communication, and intelligence units; and ground support equipment. It is estimated to cost $17.6 billion. The DarkStar was a high-altitude unmanned aerial vehicle designed to provide theater reconnaissance and surveillance to operational

commanders. It consisted of an air vehicle piloted remotely from the ground and a ground control station. Its total cost was $212 million. The SLAM-ER is a Navy missile that will be used on aircraft carriers and launched from an F/A-18 aircraft to make precision strikes against land targets. The expanded response missile, a follow-on to the original SLAM missile, is designed to have a longer range, increased probability of destroying targets, increased system lethality, and improved guidance and navigation. Its estimated program cost is $525 million. The Air Force's F-22 aircraft is an air superiority fighter designed to succeed the F-15. It is designed with low radar observability, supersonic cruise capability, and sophisticated avionics. Its estimated program cost is $62.5 billion. Cost figures for the above programs are represented in then year dollars.

In analyzing the reasons why differences existed between DOD's testing practices and those of the firms we visited, we drew on both work done for this report and previous work done on best practices. In particular, the information we present on the factors that comprise the business cases—or justification for new program or product developments—draws heavily on our previous reports and testimonies we have issued. These can be found in the list of related GAO products at the end of this report.

We conducted our review from March 1999 through June 2000 in accordance with generally accepted government auditing standards.

Problems Found Late in Development Signal Weaknesses in Testing and Evaluation

Late-cycle churn is a phrase one commercial firm used to describe the scramble to fix a significant problem or flaw that is discovered late in a product's development. Usually, it is a test that reveals the problem. The "churn" refers to the additional—and unanticipated—time, money, and effort that must be invested to overcome the problem. Problems are most devastating when they delay product delivery, increase product cost, or "escape" to the customer. Most of the commercial firms we visited had experienced such problems on earlier products but found ways to avoid them on more recent products. They view late surprises in testing as symptoms that the testing and evaluation for a product was not planned well or executed properly.

The discovery of problems in testing conducted late in development is a fairly common occurrence on DOD programs, as is the attendant late-cycle churn. Often, tests of a full system, such as launching a missile or flying an aircraft, become the vehicles for discovering problems that could have been found out earlier and corrected less expensively. For example, several failures in flight tests of the THAAD system were traced to problems that could have been revealed in ground testing.[1] When significant problems are revealed late in a weapon system's development, the reaction—or churn—can take several forms: extending schedules to increase the investment in more prototypes and testing, terminating the program, or redesigning and modifying weapons that have already made it to the field. These outcomes have broader implications for DOD's overall modernization as well, because the additional investment that is needed to correct the problems of one program is often made by cutting the funding of other programs.

Problems Revealed Late in Testing Are a Major Source of Disruption in DOD Programs

Over the years, we have reported numerous instances in which weapon system problems were discovered late in the development cycle. Differences in the type of weapon system, the complexity of the design, the respective military service, or the acquisition strategy being followed have not mattered. The corrective action most often taken was to restructure the development program so that the weapons could be redesigned and re-tested before production or to redesign and retrofit weapons in production. Rarely did a poor test result lead to program termination. The

[1]The THAAD missile system is currently in the engineering, manufacturing and development phase. The problems referred to occurred in an attempt to provide an early operational system that could be fielded.

following are examples of weapon systems that have enforced testing problems late in development.

C-17 Globemaster II Aircraft	Family of Medium Tactical Vehicles
ALQ-135 Radar Jammer	ALR-67 Radar Warning Receiver
V-22 Osprey Aircraft	Sensor Fused Weapon
B-1B Lancer Bomber	Pioneer Unmanned Aerial Vehicle
B-2 Spirit Bomber	Pioneer Unmanned Aerial Vehicle
Tacit Rainbow Missile	M-1 ABRAMS Tank
F-18E/F Hornet Aircraft	Sergeant York Artillery Gun
F-14D Tomcat Aircraft	Standoff Land Attack Missile
Rolling Airframe Missile	Aquila Remotely Piloted Vehicle
High Mobility Trailers	DarkStar Unmanned Aerial Vehicle
F-22 Raptor Aircraft	Theater High Altitude Area Defense Missile
Airborne Self-Protection Jammer	

Source: GAO.

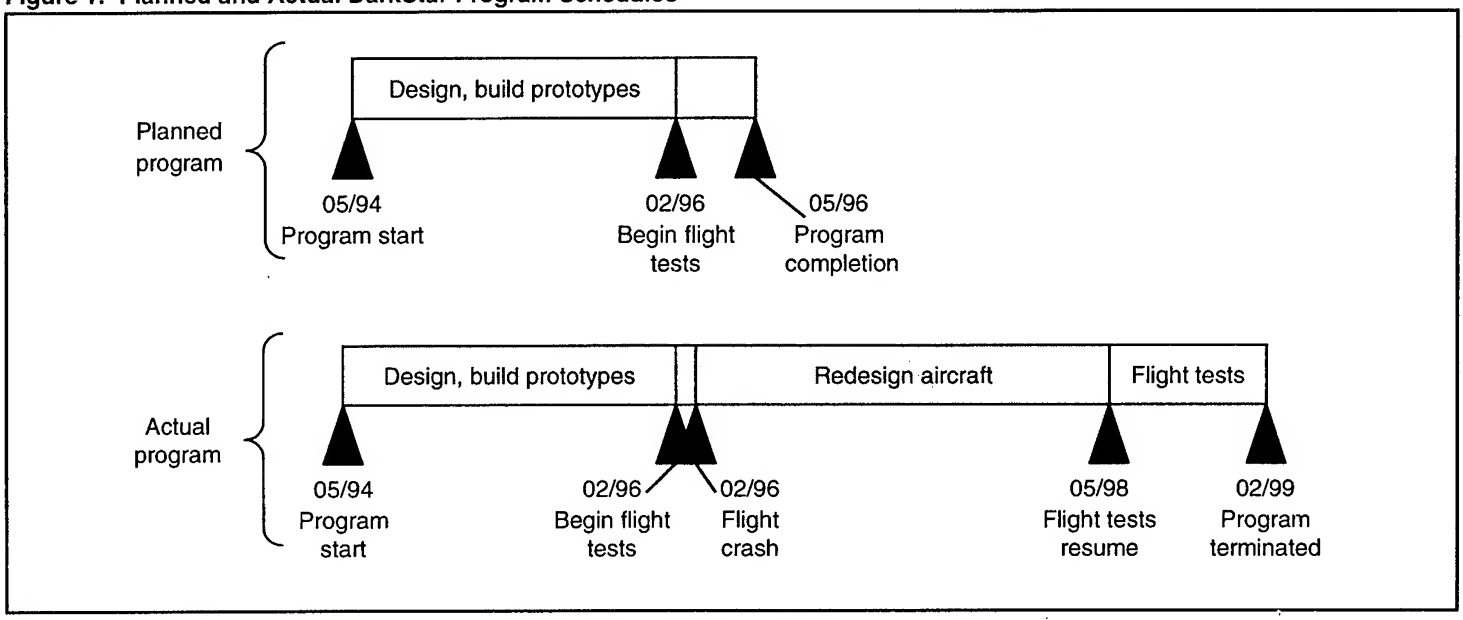
Described below are four programs that have recently experienced late-cycle churn as a result of unexpected test results late in development.

DarkStar Unmanned Aerial Vehicle

The DarkStar development program was structured to demonstrate the military utility of the unmanned aircraft. As an Advanced Concept Technology Demonstration, the DarkStar's design was to rely on mature or off-the-shelf technologies.[2] Originally, DOD planned to develop, test, and evaluate the DarkStar in 2 years. Near the end of the 2-year schedule, the aircraft crashed in its second flight test. The ensuing redesign efforts to solve the problems caused costs and schedule to double. After over 4 years of development, the program was terminated. The DarkStar's planned and actual development schedules are shown in figure 1.

[2]DOD initiated Advanced Concept Technology Demonstrations in 1994 to help expedite the transition of mature technologies from the developers to the warfighters. The purpose of such demonstration projects is to assess the military use of a capability, such as a weapon, that is comprised of mature technologies.

Figure 1: Planned and Actual DarkStar Program Schedules



Source: GAO's analysis of DOD data.

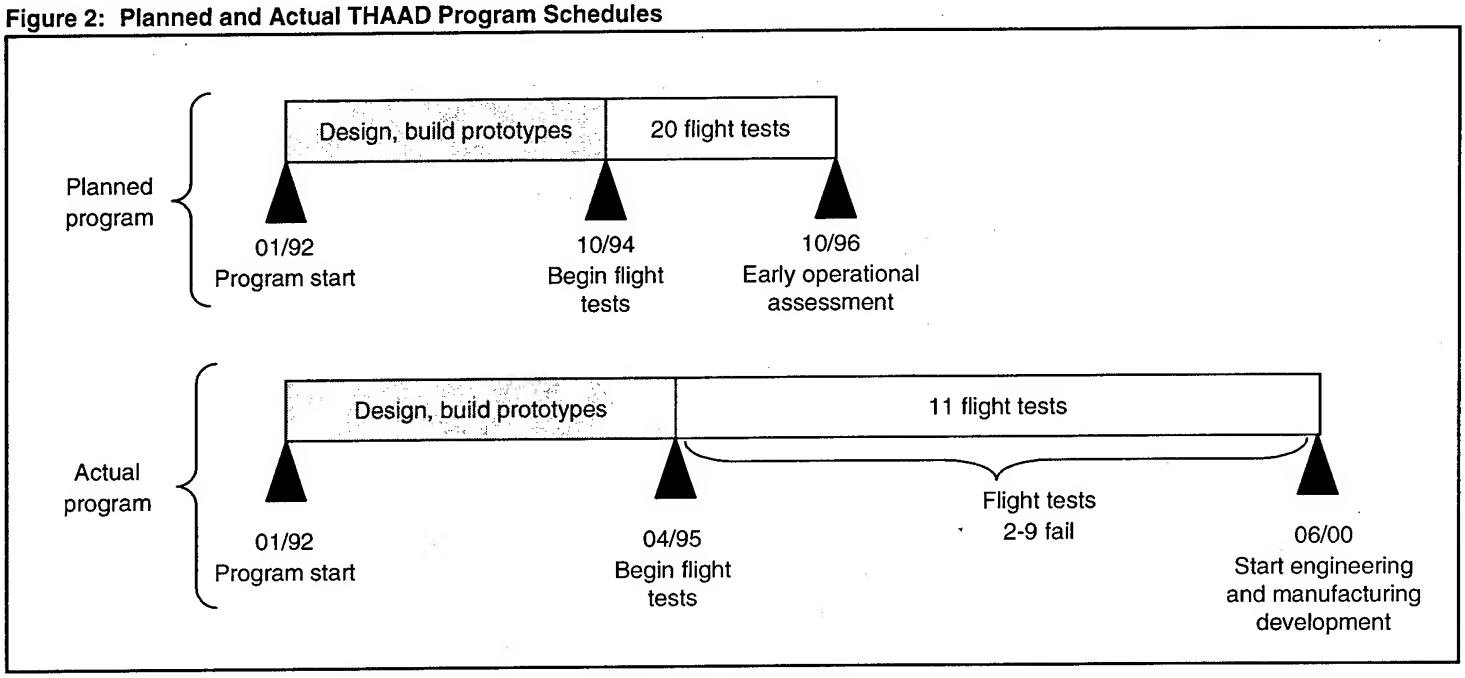
As a result of the crash, the program was extended while the contractor made significant design changes and modifications to the remaining air vehicles. Significant improvements were made in modeling and simulation, component qualification and airworthiness testing. At the end of this 2-year modification effort, the second air vehicle was flight tested, but it exhibited design flaws in the fuel subsystem. The third and fourth air vehicles incorporated design changes that resolved these problems. By that time, the program's cost had increased from $106 million to $212 million and its schedule had grown from 2 years to 4 years and 9 months. In February 1999, the DarkStar program was canceled; its termination was due to a lack of available funding for it and another unmanned aerial vehicle program. According to program officials, the later aircraft configurations showed promise, but the program was terminated before they could test them. Thus, the main purpose of the program—to determine military utility—was never achieved.

THAAD Program

The Army had planned to develop and field an initial version of the THAAD system in just under 5 years at a cost of $2.5 billion. This initial version was to provide the Army an interim capability to intercept enemy missiles, which was to be followed by an engineering and manufacturing

development phase to field a more capable system in greater numbers. As a result of problems discovered in flight testing, however, the initial version took over 8 years to develop at a cost of $4.2 billion. The THAAD's planned and actual development schedules are shown in figure 2.

Figure 2: Planned and Actual THAAD Program Schedules



Source: GAO's analysis of DOD data.

Once flight testing began in 1995, the missile experienced numerous problems. The first flight tested only the propulsion system and missile functions such as booster performance and interceptor launch. In the next eight flight tests, the THAAD missile experienced a variety of failures. Problems revealed in these tests included software errors, booster separation, seeker electronics, flight controls, electrical short circuits, foreign object damage, and loss of telemetry. These failures brought the program to the brink of cancellation in 1998. The program was subjected to four independent reviews and was significantly restructured. In the restructuring, the requirement to field an interim version of the missile was deleted. After the missile intercepted the target in the 10th and 11th flight tests, the initial version of the missile was judged successful and the program entered engineering and manufacturing development in June 2000. However, this phase will run longer than planned—over 7 years—with a commensurate delay in fielding the final missile system. Program

officials estimate that acquisition costs—both development phases and production—have increased by over $5 billion.

Army Cargo Trailer

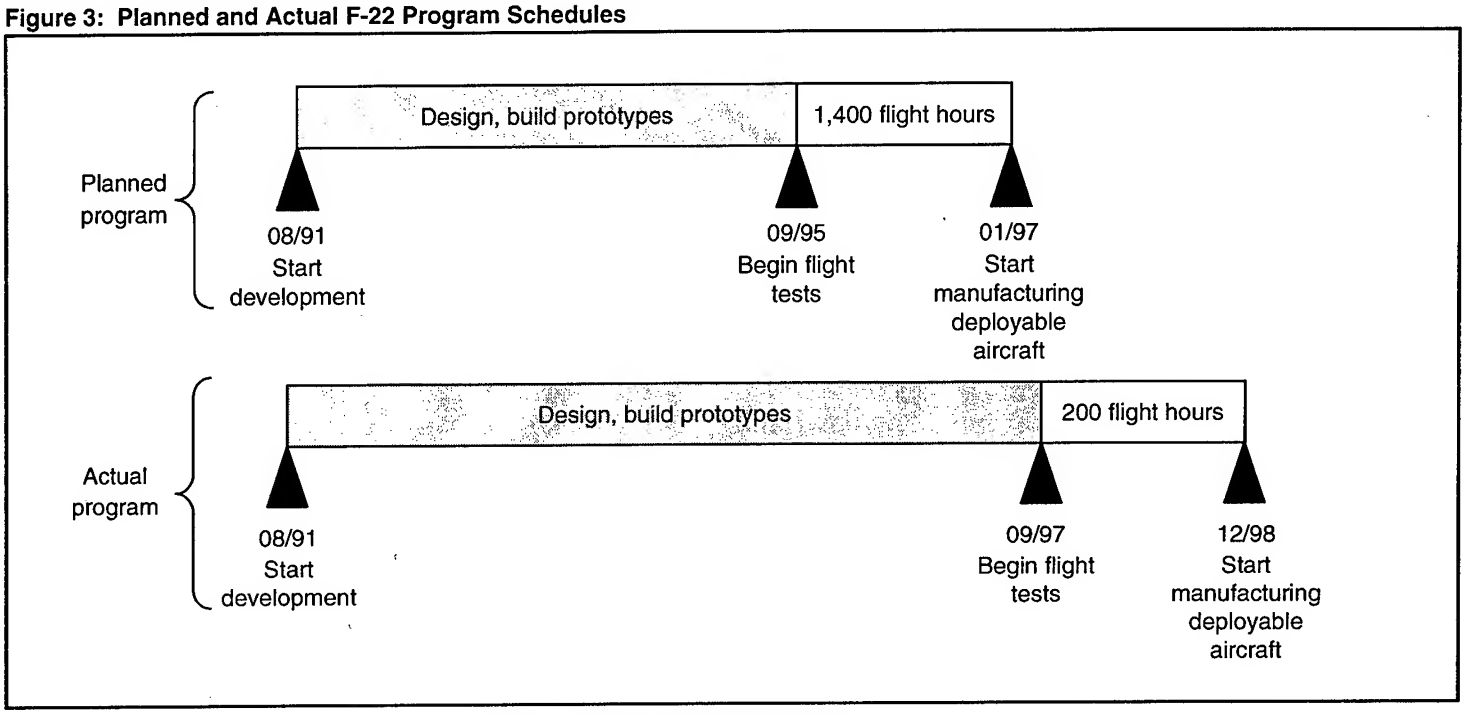
In 1993, the Army purchased about 6,700 truck trailers that cannot be used because they have serious safety problems and damage the trucks towing them. The Army entered into a 5-year production contract for the trailers without first testing the design to see if it met requirements. The Army later found that the contractor could not meet the delivery schedule, that the trailers could not pass testing, and that the trailer design would need extensive modifications. Despite these performance problems, the Army accepted 740 trailers of the original design, which it placed in operational units. After numerous problems with the fielded trailers, the Army issued a safety message that required that the trailers not be used. Since that message, the Army has continued to accept the remaining trailers from the contractor but has placed them in storage. Breaking the 5-year production contract in order to redesign the trailer is expected to increase unit costs by about 50 percent. Also, the Army will pay for modifications to the trailer and trucks; these costs have not yet been disclosed.

F-22 Air Superiority Fighter

The Air Force had planned for the F-22 to spend 5.5 years in engineering and manufacturing development before manufacturing of deployable aircraft[3] began. During that time, 1,400 hours of flight testing were planned, as shown in figure 3. Over the next few years, the F-22 development schedule was extended by nearly 4 years, the start of flight testing was delayed 2 years, and only 200 hours of flight testing were accomplished before the manufacturing of deployable aircraft began.

[3]Deployable aircraft refers to aircraft that will eventually be put into the F-22 operational fleet.

Figure 3: Planned and Actual F-22 Program Schedules



Source: GAO's analysis of DOD data.

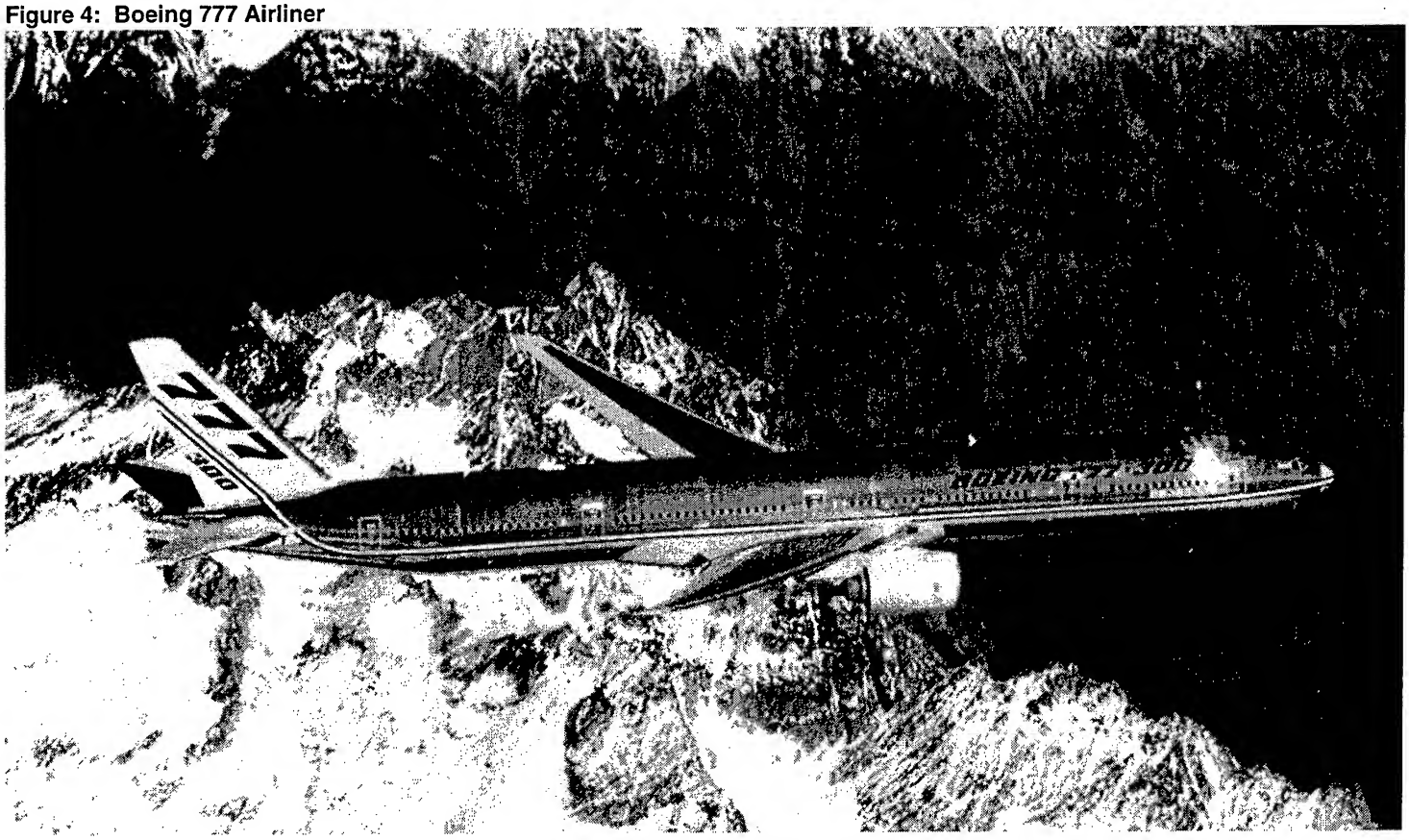
Since 1991, the F-22 aircraft has experienced (1) a number of technical and performance problems, such as software, and hardware; (2) integration problems with the communication, navigation, and identification and electronic warfare subsystems; and (3) delays in delivery of wings and aft fuselage. The effort to solve these problems has led the Air Force to extend the engineering and manufacturing development schedule from 5.5 years to about 7.5 years and increased estimated development costs from $15.3 billion to $20.4 billion—a cost cap mandated by Congress. Flight testing has also been delayed and significantly reduced in scope. The F-22 was originally planned to undergo 5,191 hours of flight testing, 1,400 of which—27 percent—were to be done by the time manufacture of deployable aircraft commenced. Current plans call for 3,757 total flight test hours. Only 200—5 percent—were completed by the time manufacturing began. The F-22 has experienced some late-cycle churn as evidenced by cost growth, schedule delays, and performance problems. The potential for performance problems in the future is significant, given that the flight testing done to date has not included all of the F-22's sophisticated subsystems (e.g., its advanced avionics). The low-rate initial production decision is currently scheduled for December 2000.

Testing and Evaluation Helps Leading Commercial Firms Avoid Late-Cycle Churn

The leading commercial firms we visited have found ways to employ testing in a way that avoided late-cycle churn yet enabled them to efficiently yield products in less time, with higher quality, and at a lower cost. Generally, these practices were prompted by problems—and late-cycle churn—encountered on earlier products. Both Boeing and Intel were hurt by new products in which testing found significant problems late in development or in production that may have been preventable. Boeing absorbed cost increases in one line of aircraft and delivered it late to the first customer; Intel had to replace more than a million flawed microprocessors from customers. On subsequent products, these firms were able to minimize such problems by changing their approach to testing and evaluation and were able to deliver more sophisticated products on time, within budget, and with high quality.

Boeing encountered significant difficulties late in the development of its 747-400 airliner, which delayed its delivery to the customer and increased costs. When the 747-400 was delivered to United Airlines in 1990, Boeing had to assign 300 engineers to solve problems testing had not revealed earlier. The resulting delivery delays and initial service problems irritated the customer and embarrassed Boeing. Boeing officials stated that this experience prompted the company to alter its test approach on subsequent aircraft, culminating with the 777-200 program of the mid-1990s. According to company officials, the 777-200 testing was the most extensive conducted on any Boeing commercial aircraft. As a result, Boeing delivered a Federal Aviation Administration-certified, service-ready 777-200 aircraft at initial delivery and reduced change, error, and rework by more than 60 percent.

Figure 4: Boeing 777 Airliner



Testing and evaluation on the 777 enabled the airliner to avoid problems experienced with previous airliners.

Source: Boeing.

A hallmark of the 777-200's success was the extended-range twin engine certification for transoceanic flight it received from the Federal Aviation Administration on the first aircraft. This certification is significant because it normally takes about 2 years of actual operational service before the Federal Aviation Administration grants extended range certification. In the case of the 777-200, the testing and evaluation effort provided enough confidence in the aircraft's performance to forego the operational service requirement.

Intel has also employed testing to avoid late-cycle churn on its new microprocessors. According to Intel officials, the company learned this lesson the hard way—by inadvertently releasing the initial Pentium® microprocessor with defects. After the release, Intel discovered a flaw in

one of the Pentium® microprocessor's higher level mathematical functions. Using analytical techniques, Intel concluded that this flaw would not significantly affect the general public because it would occur very rarely. Intel, however, miscalculated the effect on the consumer and was forced to replace more than a million microprocessors at a cost of about $500 million. Intel underwent a significant corporate change in its test approach to ensure that bugs like this did not "escape" to the public again. As a result, the quality of subsequent microprocessors like the Pentium® Pro and Pentium® III microprocessors has significantly improved. Despite adopting a much more rigorous testing and evaluation approach, Intel did not increase the amount of time it took to develop new, more sophisticated microprocessors. In fact, Intel's rate of product release increased over time.

Employing Testing to Validate Product Knowledge Early Is a Best Practice

The leading commercial firms we visited think in terms of validating a product and using testing and evaluation as a means to that end. Validation refers to verifying knowledge that a product is maturing or working as intended.[1] Thus, the focus is on attaining the necessary knowledge rather than on which techniques are used or what events are held. While individual approaches varied, the firms we visited all used validation to ensure that their products met a basic set of standards—which we refer to as product maturity levels—at given points in time. We found three product maturity levels that commercial firms had in common: technologies and subsystems work individually, components and subsystems work together as a system in a controlled setting, and components and subsystems work together as a system in a realistic setting. The key to minimizing surprises late in development is to reach the first two levels in such a way as to limit the burden on the third level. Consequently, leading firms place a high value on conducting the right validation events at the right time, ensuring that the events produce useful results, and using the results to make the product better. These firms often find that the actual effort to validate a new product's performance—indicated by test hours, for example—exceeds the effort originally planned.

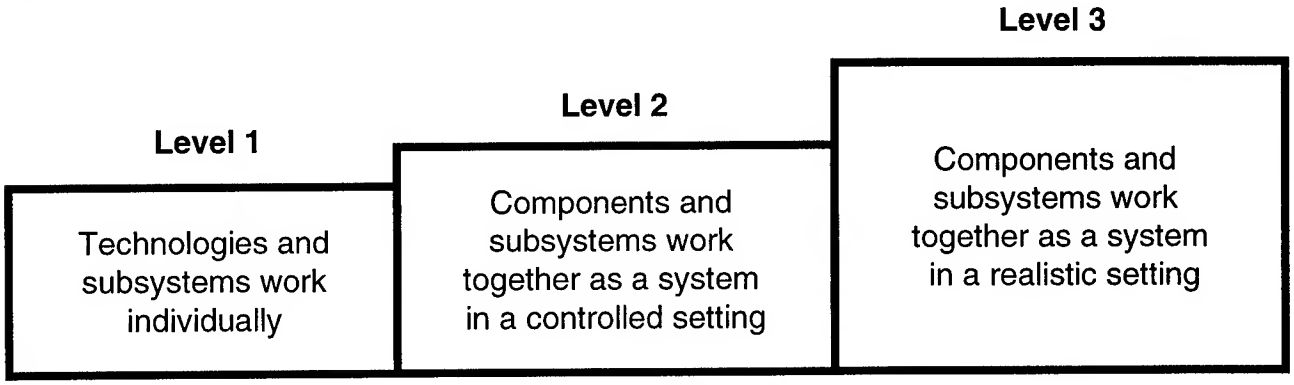
On the weapon system programs we reviewed, it was much more likely for validation of product knowledge to travel to the latter stages of development because tests were often delayed, skipped, or not conducted in a way that advanced knowledge. Consequently, the techniques and the knowledge to be gained became separated. In some cases, product knowledge was not advanced because a test was seriously flawed or because test results were not used to improve a weapon's design. Also, the amount of testing that was actually conducted on these weapons during development was often significantly less than planned. Our current and previous work on best practices has shown that because immature technologies are incorporated in weapon system designs, they are often not mature enough to be validated until late development or early production. For these reasons, the defense programs we reviewed did not validate product maturity levels as early as their commercial counterparts. Instead, product knowledge was validated later, with system level testing—such as flight testing—carrying a greater burden of discovery and at a much higher cost than found in leading commercial firms.

[1]Validation in this context differs from the standard systems engineering definition in which validation means that the system meets its real world, operational requirements.

Focusing Validation Methods on Progressive Levels of Product Maturity Reduces Late-Cycle Churn

To minimize surprises discovered through testing late in product development, leading commercial firms we visited validate a product's performance against their own standards for what knowledge should be attained at different stages in a product development. The standards—product maturity levels—we saw as being common among the firms are shown in figure 5.

Figure 5: Product Maturity Levels Commercial Firms Seek to Validate



Source: GAO.

These levels do not have to be reached in a sequential manner. For example, a firm can build a replica of a system in a laboratory, with some actual subsystems physically present and others simulated. As additional subsystems are matured, they can be brought into the laboratory to replace their simulation counterparts. However, commercial firms will not install a subsystem that has not been integrated and validated with other subsystems in the laboratory and put it on a product like an aircraft and test it in flight. The risk for surprises in a very expensive environment, such as flight testing, is too high. By focusing on the validation of product maturity levels, the firms guard against skipping or delaying a critical test or being misguided by an event that has failed to validate knowledge.

The Timing, Quality, and Utilization of Validation Events Are Critical to Reaching Product Maturity Levels

Leading commercial firms have found that if they do not validate a new product properly, they experience the same kinds of problems that DOD weapons experience late in development. These problems either disappointed customers or allowed the competition to overtake them. To reach the first two product maturity levels in such a way as to limit the burden on the third level, leading firms have learned to discipline their validation approaches to ensure that

- the right validation events occur at the right times,
- each validation event produces quality results, and
- the knowledge gained from an event is used to improve the product.

The first factor has been described as "doing the right thing," that is, having the key steps in place to validate the product. The other two factors refer to "doing the thing right." Regarding the timing of events, two features of leading commercial firms' validation efforts stand out. First, individual technologies are validated before they are included in a product's design. Second, the firms schedule challenging validation events early to expose the weaknesses in the design. AT&T refers to this as a "break it big early" philosophy. Events are comprised of a variety of tools and techniques, such as modeling and simulation, physical testing, software regression, hardware-in-the-loop, and a fault tree analysis.[2] While we found that the firms used the full range of techniques, no single technique was the most important.

The maturity of a prototype, software and hardware compatibility, realism of test conditions, and fidelity of modeling and simulation all play a role in determining the quality of validation events. It is possible to conduct a test or simulation that does not contribute worthwhile information. For example, conducting a system integration test would produce limited results if the hardware and software were incompatible. On the other hand, modeling and simulation might be performed with such a degree of fidelity that other physical tests are not needed. Thus, the same product maturity level can be reached with different tools, given the tools have the quality to produce the requisite knowledge. In other cases, modeling and simulation could not substitute for physical testing. Given that the right events have

[2]Hardware-in-the-loop is a test technique that employs system software with representative subsystem hardware to simulate a weapon system's actual operating environment. A fault tree analysis is a technique that assesses hardware safety to provide failure statistics and sensitivity analyses that indicate the possible effect of critical failures.

been held and conducted the right way, the knowledge gained must be used to improve the product. Using the knowledge depends on an accurate analysis and evaluation of an event's results and taking the time and effort to incorporate changes into the product design.

All of the firms we visited had adopted product validation approaches that have reduced the burden on system level testing late in development. These approaches feature maturing products in increasing—and well-defined—levels of maturity and testing difficult technology or design features early. Several of the firms noted that no validation approach is perfect and that validating one product the right way does not guarantee that mistakes will not be made on the next product. Examples of successful validation approaches applied by Boeing and Intel are detailed below. Details on the validation approaches of AT&T, General Electric, and DuPont are discussed in appendix I.

Boeing's Investments in the 777-200 Program Improved Both the Quality and Quantity of Validation

After experiencing late-cycle churn on the 747-400, Boeing adopted a gated product development process to validate knowledge earlier in product development. Officials referred to this as having to "move discovery to the left." Previously, Boeing engineers would continue to design an aircraft after manufacturing had begun, which limited the amount of product knowledge attained early and shifted the burden of discovery to later testing. For example, Boeing generally did not release 90 percent of engineering drawings on a new aircraft—a key indicator of design maturity—until flight testing began. On the 777-200 program, Boeing released 90 percent of the drawings about 14 months before flight testing. This accomplishment was due in part to Boeing's significant investment in design and validation processes, particularly in simulation and ground testing.

According to Boeing officials, the key to successfully integrating new 777-200 components and subsystems into a system—the second product maturity level—was Boeing's emphasis on early validation and ensuring that the critical processes, tools, and facilities were in place to perform the validation. Boeing did not allow candidate technologies, such as advanced avionics, into 777-200 components and subsystems unless they were mature. The candidate technologies were validated extensively in Boeing's laboratories through materials testing, modeling and simulation, and scale model testing—before the 777-200 program was launched. If a sufficient knowledge base could not be established or if the results indicated that a technology represented a significant risk, the technology was not included

in the design. For example, after testing aluminum lithium, a lightweight material, Boeing determined that the material was immature and thus too risky to be included on the 777-200.

To help ensure that mature test articles were available for higher level validation tests, Boeing initiated a process called proven functional operability dates. This is a process by which Boeing defines all of the functions for its components, parts, subsystems, and the dates those functions are to be ready for test. The proven functional operability dates process provides a critical path for validating the maturity of components before they are integrated into system level testing. For example, a flight control component may ultimately have to perform 10 functions. Boeing will define those functions at the start of development and lay out a schedule for validating them before flight testing. This baseline becomes the criteria for assessing not only the component's maturity but also the quality of the test event as well. If a test of the flight control is to demonstrate 7 of the 10 functions, but the article being tested has fallen behind and is capable of only 4 functions, the test has less value—it no longer validates as much knowledge. Validating the other three functions will have to be made up or the product could fall behind or build up too much risk in the later stages.

Boeing made an extensive investment in new system-level techniques to design the 777-200 and validate its maturity in a controlled setting. Specifically, a three-dimensional design tool and a systems integration laboratory improved the quality of validation—enabling component integration with little or no rework—and saved time and money. Boeing adopted a design system called computer-aided three-dimensional interactive application for the 777-200. All design drawings were done on computers, which meant that the geometric definitions of parts and tools were incorporated in a digital database. The data aided both the design of the aircraft and the design of manufacturing techniques and assembly layouts. Formerly, Boeing relied on physical mockups, which were expensive and time-consuming to construct, to design components that were difficult to accurately design on two-dimensional paper drawings. The resultant parts were inaccurate and required high rates of rework. The use of the new design tool not only facilitated final assembly, but also reduced changes, errors, and rework by more than 60 percent compared with previous practices.

Boeing also used a new technique to validate the 777-200's maturity in a controlled setting—a systems integration laboratory. The laboratory

combined actual 777-200 subsystems such as avionics, electrical system, and cockpit flight controls with simulated flight conditions. Boeing linked over 60 individual laboratories into the integration laboratory to validate the 777-200 as a system. Each simulated flight recorded measurements from all systems, giving the engineers accurate data to investigate system operation and interaction. Test problems were recorded for each flight, entered into a tracking system, and processed as a "real" airplane flight discrepancy report. In this way, Boeing used the test problems to improve the aircraft. Boeing officials informed us that the accuracy of the computer-generated information is critical to the credibility of such a laboratory and that Boeing's large base of actual data from predecessor aircraft was key to the laboratory's success. Ultimately, the laboratory added about 2,000 test hours to the 777-200 program and greatly enhanced the efficiency of subsequent flight tests. Flight testing still revealed problems, but the number of new problems was low—not significant enough to cause late-cycle churn. Boeing was able to analyze them and identify potential solutions that were validated in the system integration laboratory before being incorporated on the aircraft. The monthly test-flying hour rates of the 777-200 airplane exceeded all previous programs, yet the number of new problems found on the airplane was low.

Intel's Improved Approach to Validating the First Product Maturity Level Was the Key to Better Product Outcomes

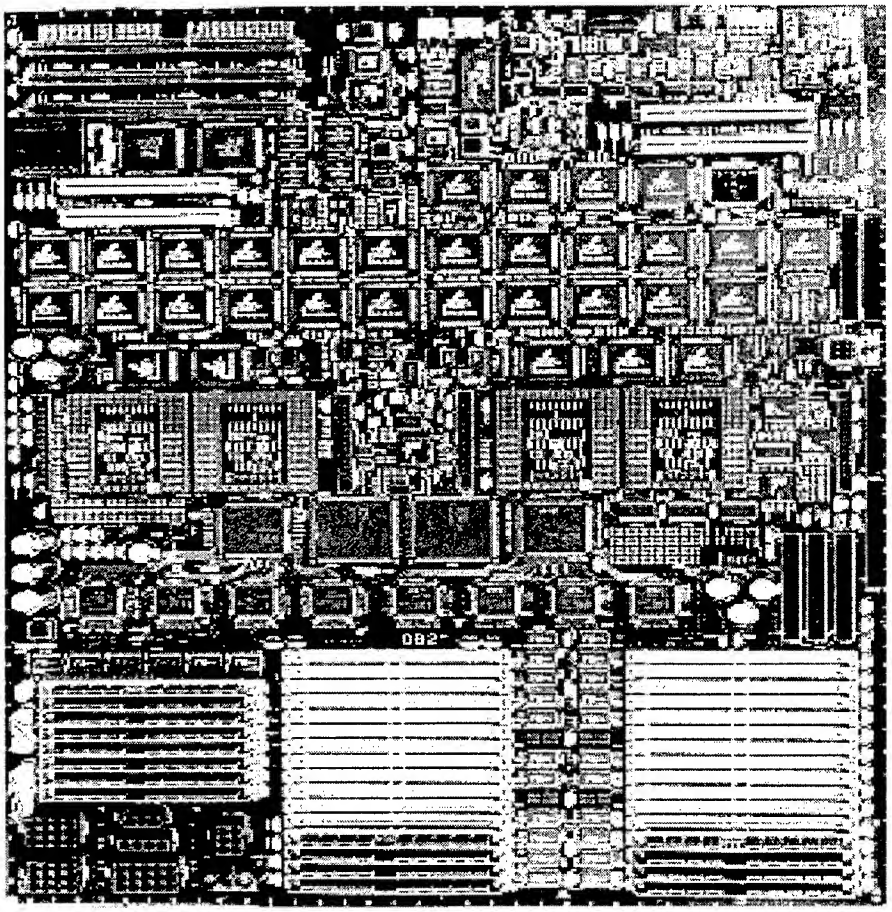
Intel's experience with the mathematics error in the first Pentium® microprocessor resulted in a significant change in Intel's approach to validating product maturity. According to Intel officials, one of the reasons the problem occurred was that testing had stopped too early. Intel instituted a validation approach to detect flaws—referred to as "bugs"—early, and more than tripled its validation staff. It adopted a three-tiered strategy to manage its new product testing: validating the microprocessor design before hardware—silicon—is made, validating prototype microprocessors in a laboratory, and validating prototype microprocessors in customers' computers. Like Boeing did on the 777-200, Intel often does more product validation than planned. According to Intel, a smaller percentage of bugs are escaping and those that do escape are less significant. In its Pentium® Pro microprocessor, Intel detected and corrected 99.7 percent of bugs before releasing the product to the public.

Like Boeing, Intel guards against ineffective tests and immature test articles, which limit product knowledge and allow bugs to go undetected. It is in the pre-silicon phase that Intel has made the most significant improvement in validation. The objective of this phase is to not only firm up the microprocessor's architecture, but ensure that the first prototype

microprocessor will be able to start and run an operating system. Intel made a heavy investment in design tools and training to limit the number of bugs designers create. It conducts extensive design validation using modeling and simulation and various software logic and architecture validation tools. Intel estimated that it finds 95 percent of all bugs during the pre-silicon phase. Intel used to proceed to silicon prototypes with less knowledge about the microprocessor. This approach relied more on finding problems in the prototypes, redesigning the microprocessor, and having additional versions of prototypes made. With the amount of validation done during the pre-silicon phase, Intel builds fewer iterations of prototypes. This, in turn, enhances productivity and reduces Intel's time to market.

Intel prototypes its first microprocessors to validate their performance in laboratory systems—the second product maturity level. This level ensures that the new microprocessor and other computer electronics integrate effectively and comply with industry standards. For its Pentium® II microprocessors, Intel built over 60 dedicated test stations with more than $100,000 of instrumentation per system. In the prototype phase, Intel also validates the compatibility of the new microprocessor with a range of peripheral computer equipment.

Figure 6: Intel's Pentium® Pro Microprocessor



Intel's investment in early validation has reduced the burden on prototype testing, such as on the Pentium® Pro microprocessor.

Source: Intel.

By the time a few customers are using the prototype microprocessor on a trial basis—the third product maturity level—product knowledge is very high. This is a marked contrast to past practices. Previously, users typically found many of the bugs. Intel now uses a variety of strategies to test a new microprocessor in realistic conditions. Computer equipment manufacturers test the microprocessor in their own products using their own methods. This strategy enables Intel to get feedback on the product's performance in a wide range of realistic environments. Intel also uses independent software and hardware vendors to test the interoperability of a wide range of peripheral equipment and software with its microprocessor. Intel developed a new user test program for the Pentium®

Pro microprocessor, selecting over 600 users to do pre-release testing of the new microprocessor. For 5 months, the users tested financial, technical, business and entertainment applications and found no new bugs.

Delayed Validation of Product Knowledge Contributes to Discovery of Problems Late in Development

In the weapon system programs we reviewed, opportunities to validate component maturity and system maturity in a controlled environment were often missed, putting a disproportionate share of validation on system testing in a realistic environment—the third product maturity level. The net effect was to attempt to reach the three levels of product maturity levels in one step in the late stages of development, which greatly increased the chance of discovering unexpected problems when the cost and schedule impact was the greatest. Testing and other techniques employed in these programs did not validate product maturity early because (1) events were skipped, postponed, or did not produce quality information or (2) information from one event was not incorporated into the product design before proceeding with the next event.

Lower Levels of Product Maturity Not Validated Before THAAD and DarkStar Flight Testing Began

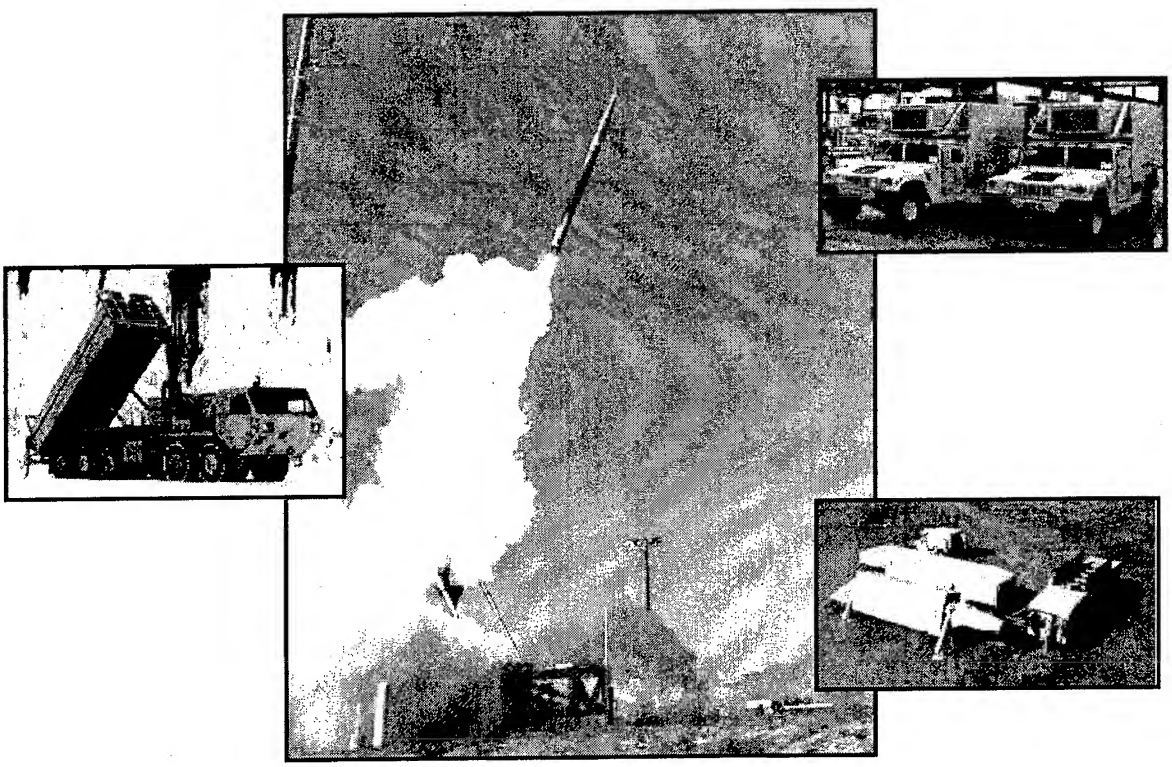
Knowledge typically gained during technology development and component and systems integration was not validated before actual flight testing began on THAAD. The planned 4-year development schedule was very tight given the THAAD's complex subsystems, several of which were unproven—a new radar, launcher system, command and control system, and an advanced seeker. The original THAAD flight test schedule left little time to develop and test technologies and subsystems.

The combination of a compressed schedule and a complex technological challenge caused much validation of the first and second product maturity levels to be deferred until the third level. Instead of a break it big early philosophy, program officials waited until flight testing to stress components and subsystems. As a result, key subsystems were not sufficiently matured for integration and flight testing. For example, the original seeker technology could not satisfy the user's needs, but tight time frames did not allow the contractor to develop a better technical solution. Although the less capable technology was chosen because it was already developed, it later proved immature. Ground testing was curtailed, and many components were shipped without thorough ground testing. Program officials acknowledged that they took many shortcuts in technology maturation, expecting to make up this knowledge during flight testing.

Validation that the THAAD components could work together as a system in the laboratory and on the ground was limited as well. While the contractor spent millions of dollars to develop an early, robust modeling and simulation tool, competing pressures to field an operational system delayed the tool, reducing its ability to help validate the integration of components and subsystems before flight testing. To stay on schedule, designers minimized the amount of test instrumentation used on the missile. Even the initial test plans did not allow sufficient time for discovery and problem resolution.

Due to time constraints, the seeker's performance was not formally validated with other subsystems before the seeker was shipped to the prime contractor for flight tests. Because the seeker developer was not allocated funding to conduct hardware-in-the-loop testing on the subsystem, the designers scavenged piece parts from around the factory to construct a test article. As a result, they did some limited subsystem testing before they shipped the seeker to the prime contractor. Software validation was also a source of problems: iterations of software were released behind schedule, lowering the quality of the tests that involved the software. Testing of avionics software was incomplete, and the corrections that were identified were not well tracked or documented. Hardware-in-the-loop testing was impaired because the hardware was not tested with the right software configuration. In addition, the missile had not been ground tested with all systems operating while subjected to thermal or vibration stresses, which it would encounter in actual use conditions.

Figure 7: The THAAD Missile System



Several THAAD flight test problems were traced to components that were not adequately ground tested.

Source: Lockheed Martin.

The limited component and integration testing placed the main burden of validation and discovery on flight testing. The failures in flight tests two through nine evidenced the amount of discovery that had to occur in the final level of product maturity. The contractor's ability to analyze why these tests failed was hampered by the elimination of earlier validation events. For example, the missile had been enclosed in a canister, leaving no hook-up points to attach instrumentation for test and measurement purposes. Post-test troubleshooting and analysis were thus greatly hampered, making it difficult to isolate problems that caused a test to fail and to correct those problems before the next test. Knowledge validation was further limited by the fact that in virtually every flight test, there was a new seeker configuration. According to the seeker contractor, the original seeker had encountered many performance and schedule problems and had been redesigned frequently. During flight testing, contractor officials stated that

they ran out of seekers and had to switch to a new seeker technology. Ironically, the contractor noted that flight test failures actually worked in its favor because the resultant delays enabled the new seeker to mature in a more disciplined manner.

According to several expert reviews from both inside and outside the Army, the causes of failures in these flights included inadequate ground testing and poor test planning. One study noted that failures were found in subsystems usually considered low risk. The failures were attributed to poor subsystem design and fabrication and inadequate ground testing. According to a 1997 Army program assessment, the physical limits of components and subsystems were discovered only by accident, when something went wrong unexpectedly. Problems were compounded by insufficient time to make corrections between flight tests. For example, the faulty software logic, which caused the failure of flight test 4, could have been discovered with pre-flight tests that are fairly standard for a system of this type. Similarly, the cause of flight test 6—seeker contamination—should have been found during subsystem qualification testing. These reviews also identified significant shortcomings in design and fabrication discipline, test planning, ground testing, and preflight review.

Like THAAD, the DarkStar's components and subsystems were not adequately validated before flight testing began. Program managers curtailed some testing earlier in the program to stay on schedule. Limited knowledge about the aircraft's performance contributed to the crash of the first test vehicle. For example, the fuel system was not sufficiently instrumented or ground tested before flight tests began. Some key sensor testing was deferred until after flight testing. Also, the contractor made extensive use of commercial components without testing or qualifying them for use on a military system.

Efforts to validate the DarkStar at the system level fell short as well. The modeling and simulation that was conducted before flight testing was not of high quality and did not have sufficient fidelity. It was cited as one of the factors that caused the crash. To save money, managers decided not to construct an "iron bird," which is a physical replica of the aircraft's hydraulics and mechanical subsystems. Normally, such a test bed is used to validate the integrated performance of these subsystems in a laboratory setting—a less complete and less sophisticated version of the 777-200 system integration laboratory. Finally, problems surfaced during the first flight test that were not fully investigated and resolved due to time constraints. For example, braking and flight dynamics problems were not

resolved prior to the next flight. After the crash, however, managers did improve modeling and simulation, component qualification, and airworthiness testing for the remaining aircraft prototypes.

Validation Approach Taken on the SLAM-ER Was Too Narrow

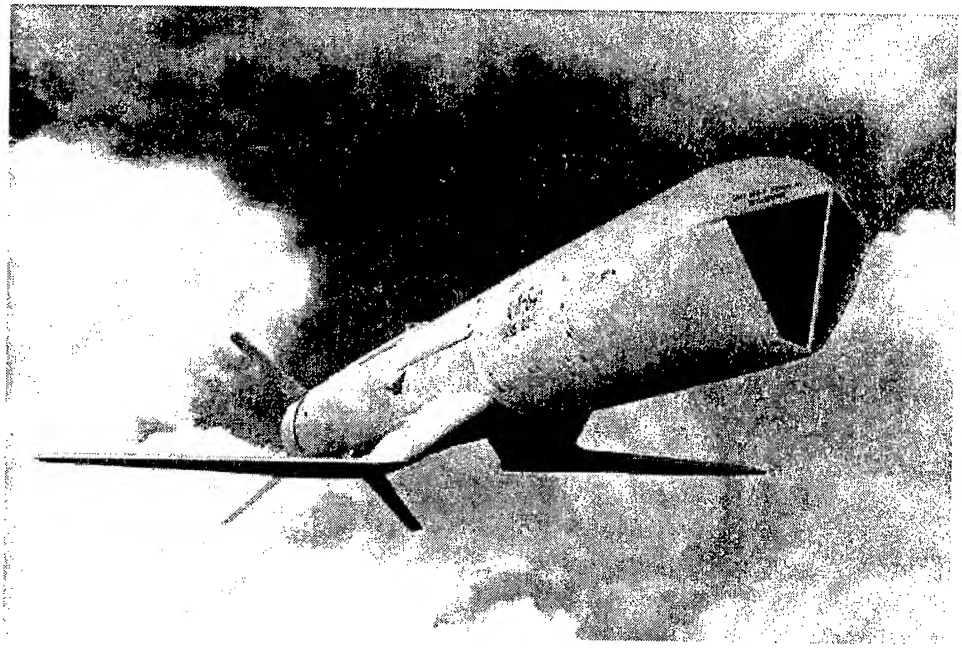
Knowledge validation before system testing in a realistic environment was also limited in the SLAM-ER program, but for different reasons. The SLAM-ER program followed a disciplined, sequential development process, with over 6,200 tests conducted in all. However, the approach was too narrowly defined and ignored problems experienced by the original SLAM missile. Also, the conditions for some system level tests were not realistic, which lowered the value of the information gained and masked some limitations. These limitations became apparent in operational testing when the missile failed to perform its mission under realistic conditions while in the hands of the customer, that is, the intended user.

According to the program manager, the SLAM-ER went through a disciplined, stair-stepped testing and evaluation process before flight testing. In its early phase, program officials identified the high-risk technologies, such as an advanced wing deployment design. The program manager used experts in national test facilities to test and refine the design. For example, experts at the Johns Hopkins University's Applied Physics Laboratory conducted extensive modeling and simulation on the wing deployment characteristics. In addition, the program used wind tunnel tests to further refine the wing design. To reduce development risks, the program manager required that all software be built and tested incrementally and reviewed by an independent team. The program also benefited from a strong corporate knowledge base; most of the management and staff had worked on the original SLAM program. All of these factors validated the maturity of SLAM-ER components and subsystems early, according to the program manager.

The program manager stated that the SLAM-ER underwent more than 3,490 subsystem and component hardware tests and over 1,800 digital software simulations. At the system integration level, the missile underwent a variety of tests, including structural tests, separation tests, hardware-in-the-loop tests, and fit checks, to ensure compatibility. Officials also conducted 12 system-level ground tests prior to flight testing. According to the program manager, developmental test results and early flight test results were promising. During initial flight tests, in which the missile performed all functions except actual firing, the missile's ability to

find targets appeared excellent. On the basis of these results, the program manager believed the missile was ready to be tested by the customer.

Figure 8: The SLAM-ER Missile



Missed opportunities to find and correct problems early contributed to the SLAM-ER's failures in customer testing.

Source: DOD.

However, during operational testing by the customer, only 5 of the 11 missile firings were successful. The tests revealed problems with the seeker's ability to find and track the intended target. Many of the problems were some of the same problems that had made users reject the original SLAM missile. These problems included signal interference, poor ability to find the target, and poor image resolution, all of which impaired the missile operator's ability to see the target through the missile seeker and guide the missile to the target. These problems were not addressed in the SLAM-ER development, yet they were significant enough to preclude the missile's ability to find and track the target. They also shortened the distance from which the missile can be fired.

Other failures were attributed to subsystem problems that went unnoticed because of quality limitations in earlier testing and evaluation. For

example, hardware and software versions did not match, which degraded the maturity of test articles and made it difficult to isolate the cause of problems. In some cases, SLAM-ER developmental flight tests were designed not so much to validate product maturity as to succeed. Test pilots and maintenance crews had become expert and intimately familiar with the test missiles. Thus, they knew how to work around problems, such as when the video images on the target acquisition system froze. For example, for one test, the ground crew heated up the intended target area to help the heat-sensing seeker. In addition, developmental test conditions were carefully controlled and test articles were prepared and maintained to be in the best condition. Finally, the system configuration was not stable—it was changed even during operational tests. The cumulative effect of these conditions was to limit the knowledge gained from the tests, allowing discovery of problems by the customer's pilots and maintenance crews.

Different Incentives Make Testing a More Constructive Factor in Commercial Programs Than in Weapon System Programs

The Under Secretary of Defense for Acquisition, Technology, and Logistics, before he took office, pinpointed the following differences in commercial and DOD testing:

> In the commercial world, the reason for testing and evaluating a new item is to determine where it will not work and to continuously improve it . . . . Thus testing and evaluation is primarily for the purpose of making the best possible product, and making it as robust as possible . . . . By contrast, testing and evaluation in the Department of Defense has tended to be a final exam, or an audit, to see if a product works. Tests are not seen as a critical element in enhancing the development process; the assumption is that the product will work and it usually does. Under these conditions, the less testing the better—preferably none at all. This rather perverse use of testing causes huge cost and time increases on the defense side, since tests are postponed until the final exam and flaws are found late rather than early.[1]

We have found similar differences in testing practices. On the basis of our current and previous work on best practices,[2] we believe these differences reflect the different demands that commercial firms and DOD impose on programs. The way success and failure are defined for commercial and DOD product developments differs considerably, which creates a different set of incentives and behaviors from program managers. Leading commercial firms insist on a solid business case for starting a new product, which centers on designing and manufacturing a product that will sell well enough to make an acceptable profit. Successful management of a product hinges on identifying unknowns early and resolving them. The role of testing or validation under such a business case is constructive, for thorough and early validation helps eliminate unknowns. Accordingly, problems that are found in testing do not threaten the product, and product managers view testers as valuable contributors to the product's success. Consequently, the leading commercial firms we visited have committed to disciplined validation approaches and to the resources necessary to carry them out.

The business case for a weapon system program is different; it centers on providing a superior capability within perceived time and funding limits. Success is more influenced by the competition for funding and the quest for superior performance. Significant unknowns are accepted in the DOD environment. Delivering a product late and over cost does not necessarily threaten program success. Testing plays a less constructive role within the

[1]Defense Conversion: Transforming the Arsenal of Democracy; MIT Press, 1995.

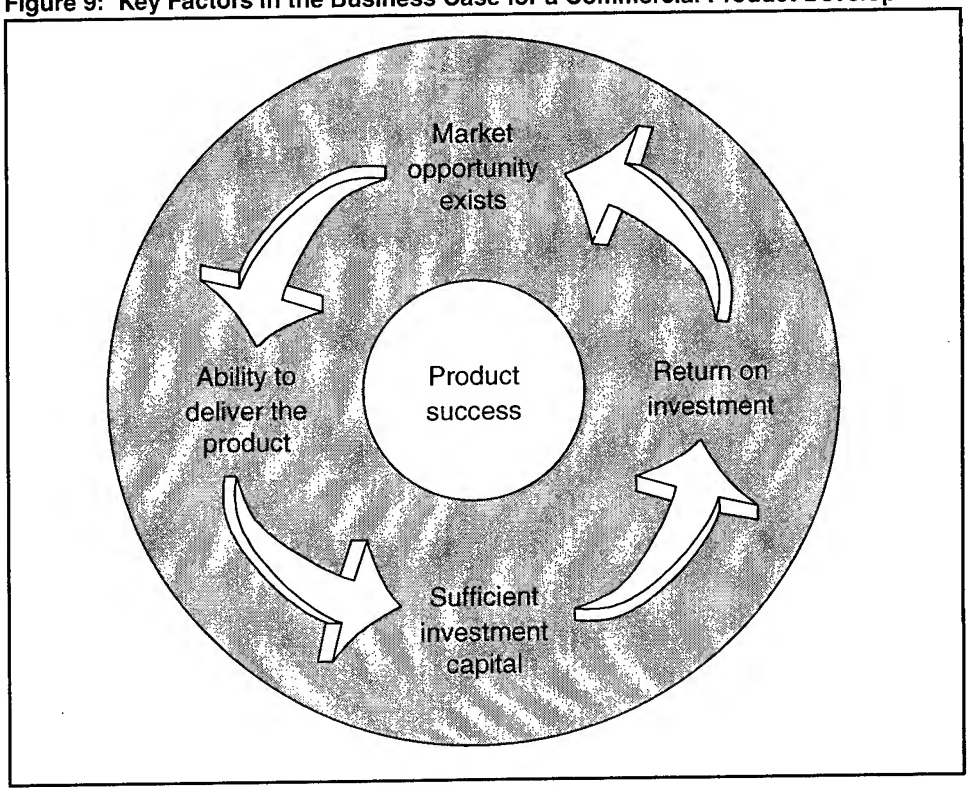
[2]See Related GAO Products.

DOD business case; a failure can jeopardize the next increment of funding and thus becomes an obstacle. Testers consequently have a more adversarial relationship with program managers. Compromises in test approaches and resources are more readily made in weapon system programs in deference to other priorities, such as keeping advertised costs low. The cumulative effect of these pressures is to defer validation of product knowledge to the end of the development phase, the conditions that lead to late-cycle churn.

Testing Is Critical to the Success of Commercial Product Developments

The main focus of a commercial product development program is to produce and sell the right product at the right time. On the basis of our current and previous work on best practices, we have identified several factors that are critical to establishing a sound business case for undertaking a new product development (see fig. 9).

Figure 9: Key Factors in the Business Case for a Commercial Product Development



Source: GAO.

While leading commercial firms have their own unique processes for starting a new product development, we have found these basic factors have to be present in some form for a commercial product to be successful. If the firm does not accurately gauge the customers' needs and determine that there is a market for the potential product, the product may not sell. If the firm does not have the technology or the engineering expertise to design a product with the features that the customer wants and to bring it to market on time, a competitor may win the customer's business. Commercial firms must spend their own money on developing new products; if they do not have the financial resources to develop the product properly, they cannot go forward with it. Finally, the firm must be able to manufacture the product at a cost that will enable it to sell with a reasonable return on investment. Cost, in this sense, includes the quality of the product because the cost of warranty repairs and returns must be factored into profit calculations.

If a product's business case measures up, the company commits to the entire development of the product, including the financial investment and the support of all company organizations. On the other hand, if any of these factors get out of line, the product may not sell and the customer could walk away. In the short term, this causes a company to lose its investment and to forego profit. In the long term, it could mean that the company's reputation is damaged. This environment encourages realistic assessments of risks and costs; doing otherwise would threaten the business case and invite failure. For the same reasons, a high value is placed on having the knowledge needed for making decisions. Program managers have good reasons to want risks identified early, be intolerant of unknowns, and be conservative in their estimates.

Commercial Incentives Foster Candor and Realism in Product Validation

Once a company decides to launch a product development, strong incentives, grounded in the business case, encourage a focus on product validation to keep the program on track. To meet market demands, leading commercial companies plan around comparatively short cycle times—often less than 2 years—to complete a product's development. These short time frames make customer acceptance and return on investment close at hand. Consequently, production looms as a near-term reality that continues to influence subsequent product decisions within the framework of the business case.

To deliver the product on time, commercial firms insist on validating the maturity of technologies before they are allowed onto a new product.

Keeping technology development out of the product development reduces the scope and risk of testing and helps make delivery times predictable. For the same reasons, the commercial firms we visited emphasize the validation of product knowledge as early as possible. The corporate commitment that product developments receive from the start defuses individual test results as a threat to program support. Because test failures do not jeopardize a program, leading commercial firms can use testing to foster knowledge and improve the product. The commercial firms we visited actively encourage testers to uncover problems. Candor is rewarded by the success of the product. Problems are expected, even welcomed, in new developments; testers are responsible for finding problems as soon as possible so that they may be corrected. The earlier problems are discovered, the less expensive and easier they are to fix. Tests are not considered failures if the product has problems or cannot achieve the anticipated performance goal; they are only failures if they do not provide insights into the product's performance.

Intel accepts the fact that bugs are inevitable, no matter how well the product has been designed. However, it can no longer afford to have a microprocessor with serious bugs to escape because production and distribution rates of microprocessors have skyrocketed since the initial Pentium® microprocessor. Today, a mistake could be in the hands of millions of customers in a matter of months. Under these circumstances, if a problem like the one on the Pentium® microprocessor escaped into the public today, the financial consequences of having to replace millions of microprocessors would be far more serious.

This fact forces an aggressive approach to validation—a change Intel officials called both cultural and procedural. Recognizing the need to overcome what one official called "the human tendency to try to rationalize or ignore problems," Intel strongly encourages its validation staff to find bugs in its products. No one is ostracized for either creating or detecting a bug. As a result, Intel has been successful in getting its validation staff to actively seek out problems and communicate them to product managers in order to improve product quality. Similarly, Boeing officials characterized problems as "gems to be mined" and said they were motivated to find and resolve problems as early as possible. During the development of the 777-200, Boeing leadership stressed the need for full and open communication—there was no kill the messenger syndrome.

In the early 1990s, DuPont became increasingly concerned about the time it took to get a new product to market. It typically took up to 6 years for

DuPont to deliver new products, while its competitors delivered products in about half the time. DuPont estimated that it could lose about $100 million for every 2-year delay on a project and concluded that if it could cut cycle time in half, it could increase revenues by 40 percent. Using these analyses, DuPont revamped its product development process, which reduced the cycle time by 40 to 60 percent. Inherent in the revamped process was an increased emphasis on testing as a means to validate product knowledge. DuPont underwent a cultural change that enabled the product designers and testers to redefine the meaning of "test failure." Previously, a test failure meant that the technology or product did not achieve expectations. Now, a test is only considered a failure if it does not produce useful information. Product failures during testing are not only expected, they are designed to occur. As long as information about a product is gained that will enhance development, the test is considered a success.

Within this constructive view of testing, commercial companies involve their testers throughout the product design, development, and production processes. Companies do this by not only making testers part of the product team but by giving them an equal voice. Testers in commercial firms are typically high-performing staff with a lot of experience and thus, credibility. The testers have a say in product development decisions. For example, Intel asks its validation staff—not its designers—whether a design is worthy of proceeding to the silicon development phase. Testers work cooperatively with the design and manufacturing engineers to devise the project plan, including the types of tests required, the timing of the tests, and the duration of the tests. Test staff also provide expert advice on how a product can be designed to facilitate testing. The significant influence of test and validation staff on a commercial product is not due to their organizational position or their ability to withhold approval. Rather, they have influence because they (1) help a product succeed and (2) are credible and have earned the confidence of the product developers.

Realistic Validation Plans and Resources Help Commercial Products Succeed

Just as candor helps cast commercial testers in a constructive role, realism helps make resource and schedule estimates accurate and predictable. Consequently, leading firms commit to thorough planning and resourcing of product validation. Commercial program development teams use fact-based estimates—proven on past programs—to arrive at realistic schedules. Optimistic schedules and resource estimates invite failure because management will have to reallocate resources to a program to cover cost overruns. Also, the customer may walk away. Conversely, a

sound and well-resourced validation approach will help ensure deliver a quality product on time.

Boeing develops an aircraft development schedule using lessons learned and actual experiences from predecessor programs. Boeing officials described the testing and evaluation conducted on the 777-200 as the most comprehensive they have ever done. While better ways of validating product maturity are encouraged, optimistic departures from successful practices are not. Boeing's commitment to delivering a product to a customer is built around this validation approach. A critical part of the resource decision is the need to have continuity of technical staff and test facilities. Boeing counts on staff to carry their knowledge and experience from one program to another. Test managers are heavily involved in test planning in order to maximize knowledge on the product as early as possible and take full advantage of Boeing's test centers. Boeing has made a significant investment in its own laboratories, and these facilities can be staffed around the clock when necessary. Boeing officials observed that this flexibility is crucial when deadlines are approaching because external test facilities are rarely so accommodating.

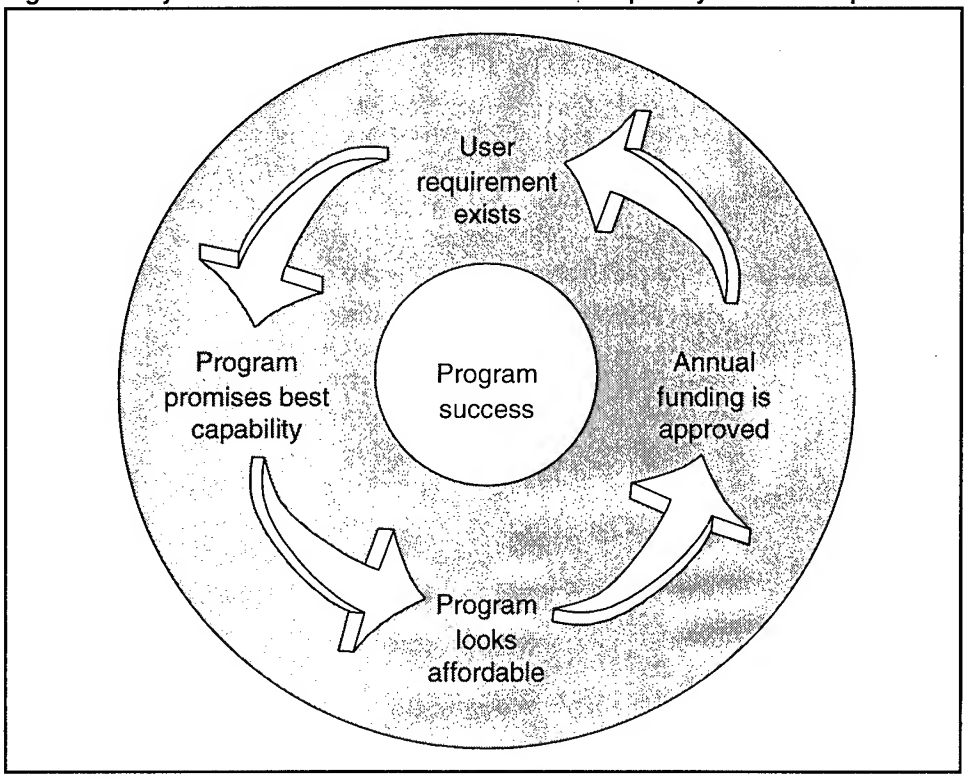
At AT&T, testing requirements are developed at the starting point of the product development schedule. Thus, curtailing or delaying tests to regain a schedule slip is inconsistent with its quality goal. Company officials stated, however, that they are not inflexible; if prior test results are convincing, AT&T can revise its test plans to delete tests that all parties agree are unnecessary. In other words, validation of maturity—not expediency—is the determining factor. Likewise, once corporate resources are allocated to the program development and testing efforts, they are made available to the development team so product success is not compromised. The resources that a product developer makes available do not define the scope of testing.

DuPont's validation process also includes rigorous test planning. Using early test results, the team makes a recommendation whether further company resources should be allocated to the program. There are strong incentives for the teams to be very informative and "honest" in their recommendations to management. Often, teams will be rewarded if they can show that resources should no longer be focused on their program.

Testing Is Perceived as Impeding the Success of Weapon System Programs

The basic management goal for a weapon system program in DOD is similar to that of a commercial product: to develop and deliver a product that meets the customer's needs. However, the pressures of successfully competing for the funds to start and sustain a weapon system program create incentives for launching programs that embody more technical unknowns and less knowledge about the performance and production risks they entail. On the basis of our present and previous work, as well as our review of outside studies, such as those sponsored by DOD, we have identified several key factors that affect the business case for starting a new weapon system program. These factors are shown in figure 10.

Figure 10: Key Factors in the Business Case for a Weapon System Development



Source: GAO.

Although DOD is attempting to create more flexibility in how technical requirements are set, a new program is encouraged to possess performance features that significantly distinguish it from other systems. The competition with other candidates for meeting the user's requirements

creates incentives for an aspiring program to include performance features and design characteristics that enable it to offer the best capability. A new program will not be approved unless its costs fall within forecasts of available funds. Because cost and schedule estimates are comparatively soft at the time, successfully competing for funds encourages the program's estimates to be squeezed into the funds available. Unlike the commercial business case, once a DOD program has been approved, it does not receive full support. The program must compete to win its next increment of funding in each budget cycle.

These pressures and incentives explain why the behavior of weapon system managers differs from commercial managers. Problems that are revealed in testing or indications that the estimates are decaying do not help sustain funding support for the program; admission that costs are likely to be higher could invite failure. Rewards for discovering and recognizing potential problems early in a DOD program are few. In contrast with leading commercial firms, not having attained knowledge—such as on the performance of a key technology—can be perceived as better than knowing the problems exist. When valid test results are not available, program sponsors can assert projected performance.

Testing Can Pose a Serious Threat to a DOD Program

Within the DOD business case for the programs we reviewed, test results tended to become scorecards that demonstrated to decisionmakers that the program was ready to proceed to the next acquisition phase or to receive the next increment of funding. As a result, testing operated under a penalty environment; if tests were not passed, the program might look less attractive and be vulnerable to funding cuts. Managers thus had incentives to postpone difficult tests and limit open communication about test results. Under these conditions, demonstrations that show enough progress to continue the program are preferred over actual tests against criteria, which can reveal shortfalls. Accordingly, DOD testers are often seen as adversaries to the program. In general, testers are often organizationally removed from the design and development effort and are viewed as outsiders. Unlike their commercial counterparts, they do not have a strong voice in early program planning decisions. As a result, their authority or influence is limited, and they are often overruled in decisions to proceed with programs despite testing weaknesses.

The role testing plays in DOD programs was analyzed in a September 1999 report from the Defense Science Board.[3] The Board concluded that the "response to perceived test failures is often inappropriate and counterproductive." Instead of using testing, especially in the early stages, as a vital learning mechanism and an opportunity to expand product knowledge, testing is often used as a basis for withholding funding, costly rescheduling, or threats of cancellation. The Board stated that distrust remains between the development and test communities, noting that some program offices have been reluctant to involve these communities early in an attempt to maintain control of the early test results. The Board also stated that testers have some reluctance to get involved early for fear of losing their independence and that this has led to polarization between the two groups when they should be united to produce a quality and robust weapon system. The Board recognized that because testers are not involved in the early stages of developing a test plan, their influence is minimized. When they are involved, they are considered disrupters rather than helpers who can anticipate problem areas and seek remedies to avoid them.

These forces were present on the THAAD program. The establishment of an early fielding requirement changed the program's priorities and became critical to the program's perceived success and funding support. The steps normally taken to validate the maturity of components, individually and integrated, before flight testing conflicted with the accelerated schedule and were curtailed. Instead, program support became equated with the results of the flight tests. When numerous test failures occurred, the program was threatened with termination. Failures, made more likely by the accelerated schedule, were at the same time less tolerable. Although the test staff raised numerous concerns about the elimination of component and ground tests, program managers who were intent on meeting the operational deadlines overruled them. In addition, contractor officials informed us that they built the first test items without including normal test instrumentation, over test staff's objection. According to the test staff, the program manager decided test instrumentation would have weighed too much, so it was deleted. The evaluation of the independent testers was highly critical of THAAD, which further distanced the test community from the program manager.

[3]Report of the Defense Science Board on Test and Evaluation, September 1999.

The test approach taken on the SLAM-ER program was consistent with incentives to avoid bad news. Its test plan did not cover resolution of serious pre-existing problems on the missile. The relationship between the testing community and the program manager eventually became a hindrance to program success. During development, testers repeatedly expressed serious concerns about the missile's capabilities and the resolution of historical problems. Funding and schedule pressures led the program manager to disregard the testers' repeated requests to investigate problems. Thus, the opportunity for the tests to make a better missile was lost. Two separate test organizations later criticized the problems the missile experienced in operational testing. When these criticisms jeopardized the SLAM-ER's approval for production, program officials believed the test staff was trying to kill the program. Eventually, a second operational test was ordered, following redesign work to correct the problems found in the first test.

The test approach taken on the DarkStar unmanned aerial vehicle was significantly compromised by cost and schedule constraints that DOD established for the program. DOD gave the program 2 years and $106 million to demonstrate military utility. Compelled by these constraints, DarkStar program managers informed us that they overruled the concerns of test officials. Little emphasis was placed on test instrumentation, software verification, automated test equipment, data analysis, and documentation. The contractor made extensive use of off-the-shelf parts, which was used as justification to skip various subsystem tests. When testers raised concerns about the need to integrate and test these items, schedule pressures prevailed over integration tests. After the first flight test, testers raised concerns about the vehicle's stability and flight worthiness, which were overridden by the managers' need to keep on schedule.

On the other hand, tester can invite an adversarial relationship with program managers. Testers can create the impression that it is more important to adhere to test regulations than it is to help make the product better. For example, one weapon system program manager informed us that the test plan was tied to an original set of requirements that the customer had since backed away from. Nonetheless, the testers insisted on testing to the requirements because that approach was in accordance with regulations and could be defended. In such cases, hollow testing—that which satisfies the requirement to hold a test but does not advance product knowledge—could actually meet some of the needs of both program manager and tester.

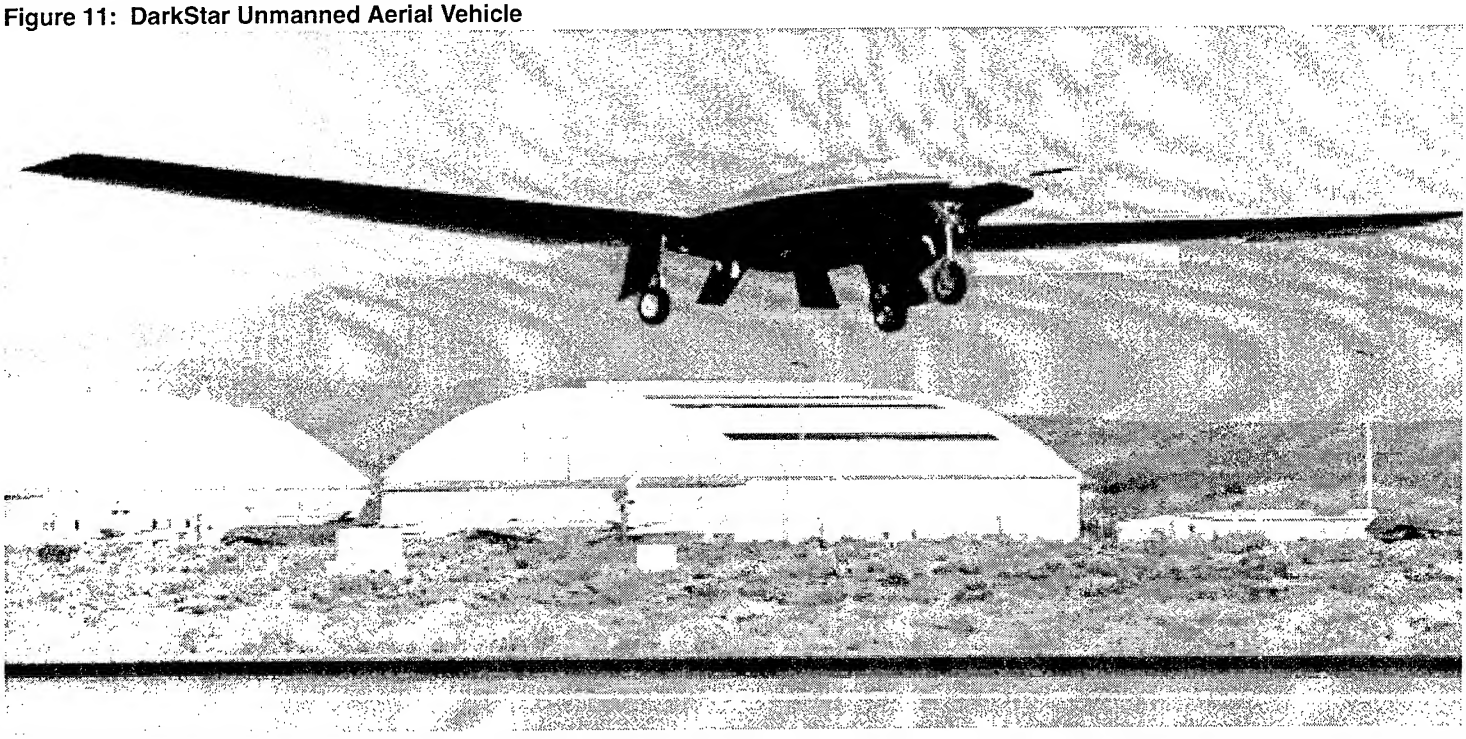
DOD Testing Impaired by Optimism and Insufficient Resources

Although DOD does extensive test and resource planning, the planning on the weapon systems we reviewed was often undercut by unrealistic assumptions. DOD's acquisition regulation 5000.2R requires formal test plans and resource estimates for every weapon system program that must be reviewed and approved by numerous organizations. This formal process does not guarantee that the program will comply with the plan or receive the resources requested or that the plan itself is realistic. On the programs we reviewed, pressures to keep schedule and cost estimates as low as possible forced managers into optimistic plans that presume success instead of anticipating problems. Test resources and schedules were assigned accordingly. The resultant test plans eventually proved unexecutable because they underestimated the complexity and the resources necessary to validate the product's maturity. Typically, the time and money allocated to testing was more a by-product than a centerpiece of the product development estimate.

The THAAD program had a requirement to develop and field a missile in 4 years that could hit and kill another incoming missile. To achieve this requirement, many unproven technologies had to come together and be proven in system tests. Yet, test plans were highly optimistic. According to program officials, the difficulty of the technology maturation process alone could not be accomplished in the time allotted. To satisfy the early fielding date, program managers opted to omit fundamental ground and subsystem tests and use flight testing to discover whether the missile design would work. When the flight tests proved unsuccessful, the early fielding date was postponed and the requirement was eventually deleted entirely.

The DarkStar test approach had similar constraints. The contractor developed a test plan that accommodated cost and schedule limits, but did not address the range of technical parameters that needed to be investigated. Problems were noted during testing, but because of schedule and cost pressures, minimal attempts were made to correct them. The safety investigation board, which investigated after the vehicle crashed, reported that "scheduling was dictated by programmatic pressures rather than sound engineering processes" and "the overriding driver repeatedly appeared to be schedule and budget." The funding and schedule constraints were imposed without considering what resources were needed to adequately mature and integrate DarkStar components into a system. Ironically, the resources to redesign and retest the system—double the original estimate—were made available only after serious problems occurred under the original plan.

Figure 11: DarkStar Unmanned Aerial Vehicle



Schedule and budget pressures significantly weakened testing and evaluation of the DarkStar.

Source: DOD.

The F-22 also constructed its test plan using optimistic assumptions. For example, program officials assumed that no hardware or software problems would be encountered during ground or taxi tests. They also assumed that one aircraft would be available for flight testing at all times and that all fights would be productive. The avionics test plan assumed a software error rate of only 15 percent, despite prior experiences of 100 percent on the B-2 and 60 percent on the C-17 aircraft. In addition, planned testing was curtailed to accommodate cost constraints on the overall program. Not only were test efforts eliminated, but the remaining tasks were often inefficiently rescheduled. Due to funding constraints, the two additional avionics test facilities recommended by an independent team were not established as intended. The first facility, an integrated hardware-in-the-loop test center, was combined with an existing test facility to save money. The second facility, an additional avionics integration laboratory, was never built due to budget limitations.

We Have Previously Recommended Ways to Make the DOD Environment More Conducive to Best Practices

We have previously reported on the need for the DOD environment and incentives to become conducive to applying best commercial practices to weapon systems.[4] We have recommended several actions that DOD could take to lessen the pressure to oversell programs and to make it more encouraging for managers to be realistic and forthcoming in assessing their programs' progress. These recommendations have implications for testing and evaluation as well and have included

- maturing new technologies ahead of and separately from weapon system programs, so that a program manager will not have to manage technology development and product development at the same time;
- redefining the launch point for a weapon system program as the point at which technology development is complete; and
- sending signals through individual program decisions that create incentives for managers to identify unknowns and ameliorate risks early in development, such as fully funding the efforts of a manager who has identified a high risk early.

DOD has agreed with these recommendations and is taking policy-level actions to adopt best commercial practices. For example, DOD is currently rewriting the directives that guide the creation and management of weapon systems—known as the 5000 series—to better separate technology development from product development. How and when these actions will be manifested on individual weapon system decisions remains to be seen.

[4] *Best Practices: Better Management of Technology Development Can Improve Weapon System Outcomes* (GAO/NSIAD-99-162, July 30, 1999) and *Best Practices: Successful Application to Weapons Acquisition Requires Changes in DOD's Environment* (GAO/NSIAD-98-56, Feb. 24, 1998).

Conclusions and Recommendations

Conclusions

We believe that late-cycle churn could be reduced and even avoided by weapon system programs if best testing and evaluation practices are applied. These practices are not tied to the use of any particular tool or technique; indeed, each firm we examined employed a unique mix of tools to test a product. Regardless of these differences, the commercial firms were alike in focusing their testing and evaluation tools on validating increasing levels of product maturity. In so doing, commercial firms guard against conducting tests that do not produce knowledge, not applying the knowledge gained from validation to improving the product, and skipping tests to stay on a schedule. Such a validation approach became central to the success of commercial product developments; schedules and resources were built around the steps needed to validate product maturity.

While DOD uses a wide variety of testing and evaluation tools, as applied, the tools often do not validate product maturity until too late, resulting in problems that require significant time and funding to correct. Several factors weaken the contribution testing and evaluation makes, particularly early in the program. These include the disruptive effects of attempting to develop technology concurrently with the product; optimistic assumptions embedded in test plans; and the fact that testing and evaluation is not viewed or funded as being central to the success of the weapon system. Under these circumstances, testing and evaluation events can become disassociated with the process of validating product maturity and illuminating areas that require more attention.

Leading commercial managers have adopted best practices out of necessity. Lessons were learned when managers ran into problems late in product development that cost the firms money, customers, or reputation. Essentially, they recognized that testing and other validation techniques, along with candor and realism, were instrumental to the success of that product. Thus, the improvement in their practices had more to do with a better appreciation for why testing is done versus how it is done. This recognition has cast validation in a different and more constructive role. With commercial testing and evaluation tools employed to advance the maturity of the product without being the basis for winning funding support, the testers were seen by product development managers as helping the product succeed. The firms have actually taken extra steps to find problems in the product because these discoveries make the product better—and not a target for criticism.

For testing and evaluation to become part of a constructive effort to validate the maturity of new weapon systems in DOD, the role it plays and the incentives under which it operates must change. Currently, testing and testers are not seen as helping the product succeed but as potential obstacles for moving forward. They become more closely linked with funding and program decisions and less likely to help the weapon system improve. Given the pressures on program managers to keep development cost and schedule estimates low, being optimistic and reluctant to report difficulties is more important to program success than planning a realistic validation effort to discover design and other problems. Attempts by decisionmakers to impose cost and schedule constraints on a program without full consideration of what is required to reach product maturity levels becomes a false discipline that can intensify pressures to defer or weaken testing, thereby increasing the potential for late cycle churn.

If DOD is successful in taking actions that respond to our previous recommendations, especially those that will reduce the pressure to oversell programs at their start, the Department will have taken a significant step toward changing what constitutes success in weapon systems and making testing and evaluation a more constructive factor in achieving success.

Recommendations

To lessen the dependence on testing late in development and foster a more constructive relationship between program managers and testers, we recommend that the Secretary of Defense instruct the managers and testers of weapon system programs to work together to define levels of product maturity that need to be validated, structure test plans around reaching increasing levels of product maturity, and orchestrate the right mix of tools to validate these levels. Acquisition strategies should then be built and funded to carry out this approach. Such a focus on attaining knowledge, represented by product maturity levels, can guard against the pressures to forego valuable tests to stay on schedule or to hold tests that do not add value to the product. This approach, which creates common ground between testers and product managers in leading commercial firms without compromising independence, still demands that the product or weapon system being matured meet the needs of the customer.

We also recommend that Secretary of Defense not let the validation of lower levels of product maturity—individual components or systems in a controlled setting—be deferred to the higher level of system testing in a realistic setting. Although the mix of testing and evaluation tools may change and the acquisition strategy may be altered during the course of a

weapon system development, the focus on attaining product maturity levels should not change. This discipline should also help guard against the practice of setting cost and schedule constraints for programs without considering the time and money it takes to sensibly validate maturity.

Finally, we recommend that the Secretary of Defense require weapon systems to demonstrate a specified level of product maturity before major programmatic approvals. In doing so, the Secretary may also need to establish interim indicators of product maturity to inform budget requests, which are made well in advance of programmatic decisions. Testing and evaluation could then be cast in a more constructive role of helping a weapon system reach these levels and would ease some of the burden currently placed on program managers to rely on judgment, rather than demonstrated product maturity, in promising success at times when major funding commitments have to be made.

Agency Comments and Our Evaluation

DOD stated that it is committed to establishing appropriate levels of product maturity, validating those levels with appropriate testing and evaluation, and providing the required mix of testing and evaluation tools necessary to validate maturity (see app. I). It agreed with two of the three recommendations but disagreed with the third recommendation. In agreeing with the recommendation that managers and testers work together to reach levels of product maturity, DOD noted that its commitment was reflected in the new 5000 series of acquisition directives and instructions, currently in draft, which is based on the concept of integrated testing and evaluation. The Department noted that testing and evaluation is the principal tool for measuring progress on weapon systems and is conducted to facilitate learning and assess technical maturity. DOD also agreed with the recommendation not to let validation of lower product maturity levels to be deferred to the higher level of system testing in a realistic setting. DOD noted that the new acquisition process model embodied in its new directives and instructions establishes entrance criteria to demonstrate that knowledge has been gained at a lower product maturity level prior to moving to the next phase of development.

The policy embodied in the new DOD 5000 series represents a potentially important step to making the acquisition process more knowledge-based. However, implementing product maturity levels on individual program level—such as when approving acquisition strategies and test plans, making funding decisions, or advancing programs through acquisition phases—will be a significant challenge. The concept of integrated testing

and evaluation is already included in the March 1996 version of DOD's directives and instructions. As it has been implemented, such integration has not included the use of product maturity levels. If decisionmakers forego the criteria and practices associated with reaching product maturity levels, as was the case on the THAAD and DarkStar programs, the new policy will be undermined and the practices that foster late-cycle churn will prevail.

DOD disagreed with our third recommendation, which originally stated that the Secretary of Defense should not allow a major test or validation event for a weapon system program to be scheduled in the same budget year as a major programmatic or funding decision. DOD stated that it could not afford to do major tests a year in advance of a decision because it would increase the costs and delay the delivery of weapon systems to the warfighter. It also expressed concerns over the potential loss of contractor engineering talent and the impact such a delay would have on the Department's goal of reducing total ownership costs and cycle time. DOD noted that rather than specify a fixed time, it must ensure that there is adequate time between the major event and the decision to evaluate the results.

Our recommendation to hold major test events in the budget year before a major program decision is scheduled was intended to make such events more constructive to furthering the development of the weapon and to lessen the threat they pose as the means decision makers use to base program and funding decisions. We did not intend—nor do we believe—that an additional calendar year would have to be inserted into program schedules. DOD's suggestion on ensuring that there be adequate time between a major test event and a major decision is worthwhile. However, as we have noted in the report, pressures still exist to make program test plans and schedules optimistic, leaving little time to resolve problems discovered in testing. Moreover, in the current budgeting process, the funds needed to execute a major program decision—such as to begin production—normally have to be requested well in advance of that decision and also in advance of key test events. DOD's suggestion does not alter these conditions or incentives. We have reworded our recommendation, dropping the language on holding test events and program decisions in different budget years, and substituting language calling for weapons to demonstrate product maturity before major programmatic approvals. We believe these changes more directly address

existing incentives and, at the same time, make the recommendation less susceptible to misinterpretation.

Validation Practices of AT&T, General Electric, and DuPont

AT&T adheres to a "break it big early" test philosophy, which is structured around quality gates. Knowledge-based exit and entrance criteria should be met before a product can move into the next quality gate. For any new product or service, AT&T works through several preliminary gates to determine its feasibility, to fully define the proposed product, and to develop a corporate commitment to it. Feasibility assessments entail tests and an analysis of the conceptual design to determine the technical maturity of the proposed product; if the maturity level is too low, the product will not proceed. Once the product has begun integration, AT&T tests the product in what it calls the "first office application." In this process, AT&T certifies product features and capabilities, conducts acceptance testing of vendor components or subsystems, and performs regression testing to identify any logic flaws. To facilitate this process, AT&T use an Integrated Test Network, which allows it to simulate products and services internally in a near-operational environment. After the successful completion of this phase, AT&T moves the product into its first field application, which is a limited trial of the product in an actual operational environment.

About 6 years ago, General Electric overhauled its product development process and developed a new, three-stage, product introduction process that stresses the criticality of early product knowledge. This new process has reduced risks and reduced development times by up to 40 percent. The first stage is a technology maturation process, which enables the company to aggressively test new technologies before committing them to a new product. Tests include modeling and simulation, characterizing the properties of new materials, feasibility, scale model, and full-size rig testing. This stage is conducted prior to a new product launch. One of the basic ground rules of General Electric's new development process is that a product must not propose a technology that has not been demonstrated through testing. Testing continues and expands to higher level assemblies in the next phase. Components are instrumented and tested in a laboratory, then integrated into subsystems and re-tested until the entire product is validated and ready for system-level testing in a controlled environment. This requires special test facilities that simulate extreme conditions for the products' use. The product then undergoes certification testing to validate performance. Once the performance is validated, the products are shipped to the buyer for additional testing on the end product.

In the early 1990s, DuPont revamped its entire product development process to reduce cycle time and become more competitive. The new process, called Product and Cycle Time Excellence, focuses on early

validation. In the first stage, technology realization, the company evaluates the commercial potential of new technologies and prepares them to be effectively used in new product developments. As part of the evaluation process, DuPont develops a matrix that identifies specific technical performance criteria and establishes feasibility points to assess progress in meeting these criteria. This is a collaborative undertaking between the business and technical communities, so when the technology is mature, it can readily transition to a new product. The second stage is the product development process. A key technique used in this stage is design of experiments. This technique accelerates the discovery process by testing several variables at one time to see how a product will react. The objective is to gain as much knowledge about the product by changing as many test variables as possible, thereby stressing the performance limits of the product. If the limits are exceeded, DuPont believes it has have gained maximum knowledge about those variables. According to DuPont officials, design of experimentation reduces the number of unnecessary tests, which reduces cost and shortens schedule. After it has successfully validated product knowledge through internal test and validation, DuPont arranges for external evaluation of the new product.

Comments From the Department of Defense

OFFICE OF THE UNDER SECRETARY OF DEFENSE

3000 DEFENSE PENTAGON
WASHINGTON, DC 20301-3000

ACQUISITION AND TECHNOLOGY

10 JUL 2000

Ms. Katherine V. Schinasi
Associate Director, Defense Acquisition Issues
National Security and International Affairs Division
U.S. General Accounting Office
Washington, DC 20548

Dear Ms. Schinasi:

This is the Department of Defense (DoD) response to the General Accounting Office (GAO) draft report, "BEST PRACTICES: A More Constructive Test Approach is Key to Better Weapon System Outcomes," dated June 19, 2000 (GAO Code 707401). The Department concurs in recommendations one and two but non-concurs in recommendation three.

Stan Z. Soloway
Deputy Under Secretary of Defense
(Acquisition Reform)

Enclosure:
As stated

GENERAL ACCOUNTING OFFICE DRAFT REPORT
DATED JUNE 19, 2000
(GAO CODE 707401)

"BEST PRACTICES: A More Constructive Test Approach Is Key to Better Weapon System Outcomes"

.....

DOD COMMENTS IN RESPONSE TO THE GAO RECOMMENDATIONS

GAO Recommendation: That the Secretary of Defense instruct the manager and testers of weapon system programs to work together to define desired levels of product maturity that need to be validated, structure test plans around reaching increasing levels of product maturity, and orchestrate the right mix of tools to validate these levels.

DoD Response: Concur. The Department is committed to establishing appropriate levels of product maturity, validating those levels with appropriate test and evaluation, and providing the required mix of test and evaluation tools necessary to validate maturity. This commitment is reflected in the current rewrite of DoD Directive 5000.1, DoD Instruction 5000.2, and DoD 5000.2-R which is based around the concept of integrated test and evaluation. The draft DoD Directive 5000.1 (currently in formal coordination) addresses integrated test and evaluation as follows:

> "Test and evaluation is the principal tool with which progress in system development is measured. … Test and evaluation is conducted to facilitate learning, assess technical maturity, facilitate integration into fielded forces, and confirm performance."

Draft DoD Instruction 5000.2, Section 2.2.1 (currently in formal coordination) states:

> "System modeling, simulation, test, and evaluation activities shall be integrated into an efficient continuum planned and executed by a test and evaluation integrated product team (T&E IPT)."

GAO Recommendation: That the Secretary of Defense not let the validation of lower levels of product maturity … be deferred to the higher level of system testing in a realistic setting.

DoD Response: Concur. The current rewrite of the Directive, Instruction, and Regulation is based on the accumulation of knowledge building towards system integration. The new acquisition process model established so-called "entrance criteria" whose purpose is to demonstrate that knowledge has been gained at a lower product maturity level prior to moving into the next phase of development.

GAO Recommendation: That the Secretary of Defense not allow a major test or validation event for a weapon system program to be scheduled in the same budget year as a major programmatic or funding decision.

DoD Response: Non-concur. We cannot afford to do major tests a year in advance of a decision. To do so would delay the delivery of needed systems to the warfighter. In addition, delaying a major programmatic or funding decision until a year after a major test event would be cost prohibitive and risk losing the contractor's engineering talent to another project within the organization or to another company altogether. Furthermore, such a delay does not support the Department's goal of reducing total ownership costs and cycle time. Rather than specify a fixed time, we must ensure that there is adequate time between the major event and the decision to adequately evaluate the results.

GAO Contacts and Staff Acknowledgments

GAO Contacts

Louis J. Rodrigues (202) 512-4841
Paul L. Francis (202) 512-2811

Acknowledgments

In addition to those named above, Jeffrey Hunter, James L. Morrison, and Rae Ann Sapp made key contributions to this report.

Related GAO Products

Defense Acquisition: Employing Best Practices Can Shape Better Weapon System Decisions (GAO/T-NSIAD-00-137, Apr. 26, 2000).

Best Practices: DOD Training Can Do More to Help Weapon System Program Implement Best Practices (GAO/NSIAD-99-206, Aug. 16, 1999).

Best Practices: Better Management of Technology Development Can Improve Weapon System Outcomes (GAO/NSIAD-99-162, July 30, 1999).

Defense Acquisitions: Best Commercial Practices Can Improve Program Outcomes (GAO/T-NSIAD-99-116, Mar. 17, 1999).

Defense Acquisition: Improved Program Outcomes Are Possible (GAO/T-NSIAD-98-123, Mar. 17, 1998).

Best Practices: DOD Can Help Suppliers Contribute More to Weapon System Programs (GAO/NSIAD-98-87, Mar. 17, 1998).

Best Practices: Successful Application to Weapon Acquisition Requires Changes in DOD's Environment (GAO/NSIAD-98-56, Feb. 24, 1998).

Major Acquisitions: Significant Changes Underway in DOD's Earned Value Management Process (GAO/NSIAD-97-108, May 5, 1997).

Best Practices: Commercial Quality Assurance Practices Offer Improvements for DOD (GAO/NSIAD-96-162, Aug. 26, 1996).

Ordering Information

The first copy of each GAO report is free. Additional copies of reports are $2 each. A check or money order should be made out to the Superintendent of Documents. VISA and MasterCard credit cards are accepted, also.

Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Orders by mail:
U.S. General Accounting Office
P.O. Box 37050
Washington, DC 20013

Orders by visiting:
Room 1100
700 4th St. NW (corner of 4th and G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders by phone:
(202) 512-6000
fax: (202) 512-6061
TDD (202) 512-2537

Each day, GAO issues a list of newly available reports and testimony. To receive facsimile copies of the daily list or any list from the past 30 days, please call (202) 512-6000 using a touchtone phone. A recorded menu will provide information on how to obtain these lists.

Orders by Internet:
For information on how to access GAO reports on the Internet, send an e-mail message with "info" in the body to:

info@www.gao.gov

or visit GAO's World Wide Web home page at:

http://www.gao.gov

To Report Fraud, Waste, or Abuse in Federal Programs

Contact one:

- **Web site: http://www.gao.gov/fraudnet/fraudnet.htm**
- **e-mail: fraudnet@gao.gov**
- **1-800-424-5454 (automated answering system)**